RICE UNIVERSITY

Practical Impact of Predictor Reliability for Personnel Selection Decisions

by

Jisoo Ock

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE

Frederick L. Oswald, Associate Professor,
Chair Psychology

Margaret E. Beier
Associate Professor, Psychology

Anton J. Villado
Assistant Professor, Psychology

HOUSTON, TEXAS
April, 2012

Abstract

Practical Impact of Predictor Reliability for Personnel Selection Decisions

by

Jisoo Ock

In personnel selection, employment tests are intended to reduce selection errors and increase mean performance. The current thesis examines the impact of psychometric properties of the predictors on *selection accuracy*, or the consistency between selection on observed scores versus true scores. Implications for validity and subsequent levels of job performance, or *prediction accuracy*, are also examined in light of common top-down personnel selection procedures. Results reflect the importance of having reliable and valid predictor measures; the work also extends ideas in the area of utility analysis.

Acknowledgements

I would like to thank my advisor, Dr. Fred Oswald, for his mentorship in completing this thesis. I would also like to thank the committee members for their valuable inputs and feedback, which have improved this paper greatly. I would like to thank my colleagues for helping me improve this paper with their honest and helpful comments. Finally, I would like to thank my friends and families for their unending support and encouragement. Without their contributions, this work would have never seen the light of the day.

Table of Contents

List of Tables

List of Figures

Practical Impact of Predictor Reliability for Personnel Selection Decisions

Organizations often base personnel selection decisions for applicants on scores from a battery of employment measures, such as measures of cognitive ability, conscientiousness, biodata, and a structured interview. The organization's determination of effectiveness of a predictor battery is usually based on factors such as (a) subgroup mean differences on the composite score, indicating the potential for adverse impact and (b) each test in the battery demonstrating job relevance and criterion-related validity for organizational outcomes of interest (e.g., performance, satisfaction, turnover). These two benchmarks are fundamental and informative in the arenas of science, practice, and litigation. The current thesis investigates how the reliability and validity of the predictors and the criterion, in light of common top-down selection procedures, produce practical consequences for which applicants get selected into an organization and which do not. Monte Carlo simulation was the tool for this investigation.

Note that it has long been acknowledged in both research and practice that the criterion of performance is best viewed as multidimensional and longitudinal in nature (Campbell, Gasser, & Oswald, 1996); therefore the validity of a selection battery will obviously depend on the substantive nature of the performance criterion (e.g., a criterion that is task-based, citizenship-based, or both; Murphy & Shiarella, 1997), differential emphasis that are placed on different domains of the performance criterion by the rater (Rotundo & Sackett, 2002), and when performance is measured over time (e.g., at the point of hire, six months later, or ten years later; Farrell & McDaniel, 2001; Keil & Cortina, 2001). A similar argument can be made for other criteria such as turnover, job satisfaction or counterproductive work behavior (Hom & Griffeth, 1991; Rusbult & Farrell, 1983; Youngblood, Mobley, & Meglino, 1983).

Previous Literatures on Utility Analysis

The *Principles for the Validation and Use of Selection Procedures* (SIOP, 2003) described the utility of a selection device as the "projected productivity gains or utility estimates for each employee and the organization due to the use of the selection procedure" (p. 48). Notably, utility estimates are open to definition here. Traditional methods of estimating selection utility (e.g., index of forecasting efficiency; coefficient of determination) rely on the criterion-related validity coefficient and the selection ratio (Schmidt, Hunter, McKenzie, & Muldrow, 1979). Three of the most popularly used utility models in selection that incorporate some of these parameters are Taylor-Russell (Taylor & Russell, 1939), Naylor-Shine (Naylor & Shine, 1965), and Brogden-Cronbach-Gleser models (Brogden, 1946, 1949; Cronbach & Gleser, 1965), each with a different meaning attached to the quality of selection. These models provide additional information on interpreting the validity coefficient in terms of its effects on selection and its implications for utility. The following sections briefly describe each model and explain how the current paper extends them.

Taylor-Russell model. In this model, Taylor and Russell defined the utility of a selection battery directly in terms of the *success ratio* that it provides. This ratio is the proportion of selected applicants who are actually successful on a job performance criterion, where *successful* is defined dichotomously (successful vs. not successful). Thus, the numerator contains the successful applicants, or the true positives, and the denominator is composed of true positives and false positives, the latter comprising applicants who were selected, but their subsequent criterion performance did not meet the threshold set by the organization.

Taylor and Russell (1939) observed that the goal of raising the success ratio is not only affected by the *validity coefficient* of the selection battery; it is also affected

by the *selection ratio* (the proportion of those selected) and the *base rate* (the proportion of applicants who would perform successfully under random selection).

It is worth explaining the base rate phenomenon. Taylor and Russell showed that utility is affected by the base rate, or the proportion of the applicant population that is competent enough to be successful on the job, prior to selection. Successful selection happens when it adds above and beyond the base rate of success by a practically useful amount. When the base rate of success is very low (near zero) or very high (near one), then it is very difficult for any selection interventions to improve upon the base rate. Conversely, a selection tool has the potential (but no guarantee) to be most useful when the base rate is equal to .50, because this creates the maximum amount of variance in the dichotomous variable of success and thus the most room for improvement.

Therefore, higher validities and lower selection ratios (near zero), and base rates closest to .50 translate into higher success ratios, and conversely, low validities (*r* near zero), higher selection ratios (near one) and base rates at the extremes (0 or 1) lead to lower success ratios.

Other things being equal, lower selection ratios are associated with higher success ratios, or the proportion of successful applicants on the job performance criterion among the hired applicants. As is evident in Figure 1, success ratio is clearly higher when the selection ratio is low ($\frac{A1}{A1+B1}$) than when the selection ratio is high ($\frac{A1+A2}{B1+B2}$). However, it should be noted that by lowering the selection ratio, the proportion of applicants who would have been successful on the job but rejected on the predictor (*false negatives*) goes up (increase from $A3$ to $A2+A3$ as selection ratio increases).

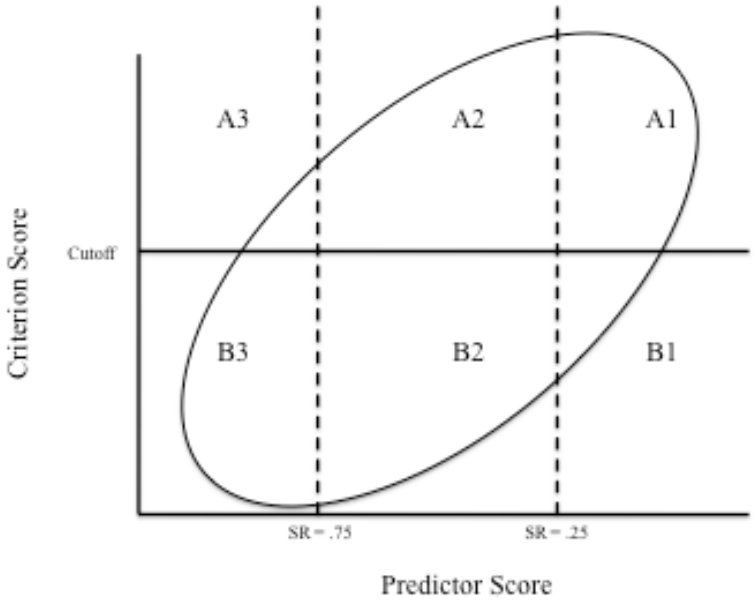

Figure 1. Effect of varying predictor cutoffs given a bivarate distribution for predictor score and criterion score. SR = .25 = predictor cutoff when the selection ratio is .25. SR = .75 = predictor cutoff when the selection ratio is .75.

In sum, the Taylor-Russell model (1939) convincingly demonstrates the effect that validity, selection ratio, and base rate jointly have on the utility of a selection tool. The model allows easy observation of the tradeoffs that are made when the selection parameters are adjusted independently or simultaneously. That said, the Taylor-Russell model makes two key assumptions that serve as its limiting factors. First, it does not explicitly account for the loss incurred to the organization by failing to hire applicants who would have been successful on the job, but were rejected based on their predictor score (false negatives). These losses are implied by, but expressed directly, in the success ratio. Second, as pointed out by Cascio and Boudreau (2008), dichotomizing the utility of the selection tool into performance success vs. performance failure makes more sense in jobs where the point of dichotomy is a critical one (e.g., at a level of minimally acceptable competence, mastery, or

asymptotic performance for most employees). However, in jobs where the relationship between the predictor and the criterion is linear, there is often a valuable difference in return between excellent performance and average performance, where all levels of performance in selected applicants yield some differential benefit to the organization. Perhaps more importantly, ineffective performance from erroneous selection may be especially debilitating for the organization, such as when job performance requires interdependent group effort, where highly ineffective performance of even one individual is damaging, or when performance contains critical components that must be performed with extremely high levels of accuracy, yet applicants vary in their success on those components.

Naylor-Shine model. In the Naylor-Shine model (1965), the value of the selection tool is defined not in terms of a success ratio but in terms of the standard deviation (*z*-score) increase in the criterion that is achieved through selection, given specified values of the criterion-related validity and selection ratio (Cascio, 1980; Hakstian, Woolley, Woolsey, & Kryger, 1991; Myors, 1993). Unlike the Taylor-Russell model, the Naylor-Shine model does not dichotomize utility of the selection tool in terms of success and failure. Rather, it assumes a continuous linear relationship between the validity and utility, where an increase in validity between *X* and *Y* leads to a proportionate criterion score increase in those selected vs. the entire applicant pool. The basic equation underlying the Naylor-Shine model is:

$$\overline{Z}_{yi} = r_{xy} \frac{\lambda_i}{\phi_i}$$

,where $\overline{Z}_{yi}$ is the standardized mean criterion score of those selected, r_{xy} is the criterion-related validity of the predictor, λ_i is the height of the normal curve at the predictor cutoff corresponding to the selection ratio, and ϕ_i is the selection ratio.

Based on this general equation, Naylor-Shine tables can be used to solve for one of the parameters, which can be used to answer several important practical questions in HR applications. For example, it may be used by organizations to determine a priori the level of parameters that are necessary to achieve a desired outcome (e.g., setting the selection ratio for achieving a desired level of increase in performance – in standardized units – given a specific validity coefficient of the predictor).

Brogden-Cronbach-Gleser model. Neither the Taylor-Russell model nor the Naylor-Shine model specifically take into account the ultimate financial gains from using a selection tool. Based on the linearity assumption between the predictor and criterion, the Brogden-Cronbach-Gleser model estimates the monetary utility of a selection tool (over choosing applicants at random) as a joint function of validity between the test scores of the predictor test *x* and the criterion performance measured in dollars, selection ratio, and estimate of the standard deviation of job performance (SD_y), or the expected value of one standard deviation increase in criterion performance, in dollars. The total expected monetary value of the selected applicants is expressed as

$$Y_s = r_{xy} SD_y \bar{z}_{x_s} + \mu_y$$

, where Y_s is the dollar value of mean criterion performance of the selected applicants, r_{xy} is the validity between the predictor test scores and performance criterion, SD_y is the standard deviation of job performance in dollars, $\bar{z}_{x_s}$ is the mean standard predictor test score of the hired applicants, and μ_y is the mean job performance of randomly selected applicants in dollars. To derive an equation that calculates the *increase* in job performance (in dollars) from using the selection process, μ_y is transposed to give

$$Y_s - \mu_y = r_{xy} SD_y \bar{z}_{x_s} .$$

The left side of the above equation is expressed as ΔU to represent change in utility, or expected financial benefits of criterion performance from using the selection process above and beyond random selection. Finally, fixed costs associated with implementing the selection process is added to the equation to calculate the marginal utility, expressed as

$$\Delta U = r_{xy} SD_y \bar{z}_{x_s} - \frac{N_a C}{N_s}$$

, where N_a is the total number of applicants, C is the cost associated with implementing the selection process, and N_s is the number of selected applicants.

This approach to estimating the utility of a selection tool is practical and easy to interpret. However, its application in the organizational literature has been with great caution because of the doubts related to the statistical assumption of the model and difficulty involved in accurately estimating the parameters, in particular the estimate of SD_y (Cronbach & Gleser, 1965; Schmidt et al., 1979). In addition, monetary gains from using the selection process is made in comparison to criterion performance from random selection, which can make the relevance of its utility estimates subject to questions, given that organizations are unlikely, if ever, to select applicants at random.

The purpose of the current thesis is to extend the past literatures on selection utility by examining the influence that measurement unreliability could have on the organization's accuracy in making selection decisions, and their subsequent implications for its productivity by bringing together the literature on classification accuracy (selection accuracy, found more in education) with a different literature on utility and validity (found more in organizational psychology). Although one might

think the loss that is incurred due to measurement unreliability might be estimated relatively quickly through psychometric corrections (e.g., meta-analysis correction formulas), the corrections are not straightforward because of multiple factors such as sampling error variance, incidental selection on a composite score, or multiple-hurdle selection. Estimation of utility, mean performance, selection accuracy, and variation in these estimates across samples turns out to be much more tractable through simulation procedures.

Classical Test Theory Approach to Measurement Reliability

A brief review of classical test theory (CTT) will be used as a means to introduce the details of the simulation that follows. CTT partitions the total variance on a variable into two independent components: true score variance and random error variance. True score variance includes variance that can be attributed to the stable characteristic being measured (e.g., cognitive ability, conscientiousness) plus any systematic biases (e.g., systematically rating an irrelevant attribute, such as the likability of an applicant; Schmidt, Viswesvaran, & Ones, 2000). Random error variance reflects unmodeled sources of variance assumed to be due to things such as idiosyncratic item-wording, differences in item forms or formats, and random fluctuations in a person's responses over time due to mood, fatigue or item-by-person interactions. Reliability is then estimated as the proportion of the total observed-score variance that is estimated as true score variance (Lord & Novick, 1968). Reliability can be estimated in various ways depending on the relevant source of measurement error one seeks to identify (Cortina, 1993), such as how highly items correlate with one another (alpha), whether the rank order of test scores can be reproduced over time (test-retest reliability) or how consistently scores are reproduced across different versions of the same test (alternate-forms reliability; Cronbach, 1947); combinations

of these factors can be incorporated into a single reliability estimate (Le, Schmidt, & Putka, 2009).

The CTT approach to measurement reliability is not without limitations. Specifically, it assumes that all unreliability is random. However, it is reasonable to assume that in addition to the random measurement error, systematic bias unrelated to the construct of interest can also distort true scores. Common method variance, such as a general liking or disliking of an employee influencing performance ratings by that employee's supervisor, regardless of the performance dimension being rated, would be one example (Cardy & Dobbins, 1986; Schmitt, Pulakos, Nason, & Whitney, 1996). As an extension of CTT, generalizability theory uses analysis of variance (ANOVA) methods to partition error variance into sources of variance due to systematic error and unsystematic error (i.e., error that is identifiable vs. error that is random, respectively; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Although generalizability theory might be viewed as an advantageous extension of CTT, it contains its own theoretical and practical estimation challenges (Brennan, 2001; Gresham, 2003), and the available data often do not allow for modeling of systematic sources of error, despite the knowledge that these sources likely exist. For example, it is unrealistic to expect a study design to incorporate comprehensive conditions under which variance attributed to the relevant measurement facet (source of variance that is of relevance in the estimate of reliability) can be generalized to the universe of generalization (Brennan, 2001; Van Iddekinge & Ployhart, 2008). In other words, the amount of variance attributed to the measurement facet may be sample specific.

The reliability of test scores is never perfect, and therefore there is always the presence of random error variance, meaning that an applicant's observed test score

will randomly deviate (be higher or lower) from his/her corresponding true score on the focal construct being measured, with the expected deviation being greater to the extent that the measure is unreliable. The standard deviation of error variance across all individuals is called the standard error of measurement (SEM), or $\sigma\sqrt{1-r_{xx}}$, where r_{xx} is the reliability estimate (thus, higher reliability means less error). The SEM indicates the amount of error to be expected across all scores, although of course, the exact amount of error for a specific person's score is unpredictable. This is why we can correct observed criterion-related validities for the measurement error variance in a test, as is commonly done in meta-analysis; however, we can never correct individual test scores for their associated errors.

Just as individual test scores cannot be corrected for measurement error, an organization cannot correct individual selection decisions made from error-laden test scores (otherwise they would). Measurement error variance thus translates into inevitable errors in applicant selection decisions – though hopefully with fewer mistakes than relying solely on human judgment. The question at hand is about the practical impact of measurement unreliability in personnel selection, and more specifically, how selection errors affect performance errors as a function of measurement error variance and the selection procedure. In addition, the applicant sample size influences how variable these effects tend to be from sample to sample.

The importance of reliable predictor measures for making personnel selection decisions is in line with recent literature indicating that the benchmark of .70 for a coefficient of reliability is probably too low for many selection purposes, and the attribution of this benchmark to Cronbach turns out to be incorrect (Lance, Butts, & Michels, 2006). On the other hand, selection errors made on the basis of a battery of measures demonstrating even a moderate level of psychometric reliability and validity

can lead to high reliability for the composite and meaningful improvements in selection. As large as some of these selection errors may be, they are likely to be less frequent than when conducting selection on the basis of HR selection managers' intuitions, fine-tuned as they might be, because ratings likely fluctuate with respect to a specific manager and/or point in time due to variables ranging from ratings scale formats, rater capacities (e.g., cognitive limitations of the rater), and rater goals (e.g., strategic decisions of the rater) (Bretz, Milkovich, & Read, 1992; Landy & Farr, 1980; Murphy, Cleveland, Skattebo, & Kinney, 2004; Saal, Downey, & Lahey, 1980).

Clinical vs. actuarial judgment. Research has consistently demonstrated that actuarial models (e.g., regression or other reasonable mechanical combinations of data) tend consistently to outperform expert judgment in the prediction of human behavior (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Highhouse, 2008). For example, Grove, Zald, Lebow, Snitz, and Nelson (2000) showed in their meta-analysis that actuarial prediction tended to produce more accurate or equally accurate prediction compared to expert judgment. Specifically, in 47% ($N = 63$) of the studies, actuarial prediction produced more accurate prediction whereas only 6% ($N = 8$) of the studies showed more accurate outcomes for expert judgment. The general advantage for actuarial prediction was consistent regardless of the field of study (general medicine, mental health, personality, and education and training), or type of judges (medically trained judges vs. psychologists), or judges' experience (inexperienced vs. seasoned). Despite the intuition that some experts are better than others in their ability to predict future behaviors (Highhouse, 2008), empirical evidence has shown otherwise. For example, Pulakos, Schmitt, Whitney, and Smith (1996) showed that even for structured interview, which curbs the influence that individual interviewer's intuitions might exert on the procedure given the formal

nature of the application and rating process (Campion, Pursell, & Brown, 1988), the differences are likely to be due to chance rather than any real effect, considering the variance in the validity of the individual interviewers' ratings. In addition, Conway, Jako, and Goodman (1995) suggested that because interrater reliability in unstructured interview is so low, its ratings could not account for more than 10% of the variance in the job performance criterion. In support of their findings, Cortina, Goldstein, Payne, Davison, and Gilliand's (2000) meta-analysis showed that unstructured interview rarely contributed to the prediction of job performance above and beyond the variance explained by cognitive ability and conscientiousness (change in R^2 ranging from .01 to .02). All things considered, expert judgment should be used to inform actuarial models but it generally does not improve up on them. However, this is not to say that all testing is mechanical. Rather, it requires careful considerations regarding its ability to predict future human behaviors in the development process, which includes the extent to which the measurements are reliable.

Previous literature on the effect of measurement unreliability on classification accuracy. Using analytic or simulation approaches, several prior studies have examined the effect of measurement reliability on *selection accuracy*, which refers to the extent that the same classification decision is made based on true score and observed score based on single tests (e.g., Bradlow & Wainer, 1998; Rudner, 2001; Subkoviak, 1976; Swaminathan, Hambleton, & Algina, 1974). However, not enough attention has been paid to situations where the selection decision is based on combining different selection measures.

As an extension to the previous literature, Millsap and Kwok (2004) used latent modeling approach to demonstrate the effect of measurement reliability on classification accuracy under various selection situations. In their simulation, they

assumed hypothetical situations where selection is based on a composite score of p measured variables. They then compared the difference in classification accuracy in cases where the sum of the error variance (θ; proportion of score variance that is not explained by the focal construct) associated with the p measured variables was high ($\theta = 3.32$ when $p = 4$; $\theta = 13.67$ when $p = 16$) or low ($\theta = 1.41$ when $p = 4$; $\theta = 3.19$ when $p = 16$). In their study, θ values were chosen to achieve desired levels of correlation between the observed and true factor scores, which may not necessarily reflect the conditions that are expected in actual selection contexts. As expected, the correlation between the observed scores and true scores differed as a function of the difference in the sum of the error variance. When $p = 4$, the correlation was .77 with high error variance; correlation increased to .88 when the sum of the error variance was low. When $p = 16$, correlation was .92 with high error variance; correlation again increased to $p = .98$ with low error variance. Correlations were higher with more predictor variables, confirming the earlier findings that longer composites tend to be more reliable than shorter composites or each constituents of the composite (Hambleton & Slater, 1997). Expectedly, selection accuracy was higher when the sum of the error variance was low for both $p = 4$ or 16.

Douglas and Mislevy's (2010) also estimated the rate of classification accuracy based on multiple measures. In their simulation study, Douglas and Mislevy generated a standard multivariate normal data set consisting of true scores for five arbitrary tests with the correlation set to $r = .60$ across all tests. For each true score, three observed scores were generated with a standard deviation equal to the SEM, where $r_{xx} = .90$ for all tests ($1\sqrt{1 - .90} = .32$). Of the three scores, the highest observed score was used to match its respective true score to determine classification accuracy. Then, a cutoff criterion was imposed on the true and observed scores, such

that the top 70% of scores (or cut score equal to -.525) are selected (or for a compensatory model, an average score of -.525 or better). Classification accuracy was then determined as the rate at which the same classification decision (selected vs. not selected) was made based on the true score and observed score.

The authors' results showed that for a compensatory model with unit-weighted predictors, the addition of more predictors resulted in higher classification accuracy compared to using a single predictor (91.41% for one test to 95.55% for five tests). However, the generalizability of their simulation study is limited in that they examined a hypothetical situation where the reliability coefficient for the predictors was uniformly high (r_{xx} = .90), generally higher than what is found for many tests used in personnel selection. In addition, Douglas and Mislevy only considered classification accuracy based on measurement error variance in the predictor battery (i.e., comparing observed scores vs. true scores), whereas the current work also considers the effects of selection accuracy on prediction accuracy.

The Current Simulations

As useful as predictor batteries may be for selection, they can never be expected to perfectly correlate with any organizational outcome of interest. Besides measurement error variance that exists on both sides of the prediction equation, unpredictable variance remains at and after the point of selection. Important influences in the workplace, such as individual differences in exposure to job-specific training, exposure to experienced team members and effective supervisors, predict organizational outcomes. Yet, applicant test scores do not necessarily predict these factors. Consequently, even applicants selected by the organization, who have satisfactory observed predictor scores and satisfactory true predictor scores, may not have correspondingly high scores on the criteria of ultimate interest to the hiring

organization. More problematic might be that the organization views its method of personnel selection as leading to choosing the most qualified applicants and yet error variance in the predictor, along with modest criterion-related validity, may lead to selecting applicants whose predicted performance is lower than what might be expected, which might even be lower than an acceptable minimum. Again, these sorts of errors are unavoidable in the real world, where only observed scores are known, not true scores.

The goal of current simulations is not to examine the mere existence of these errors but rather the extent and practical implications of them for personnel selection. Specifically, simulations examine selection accuracy when selection is based on the observed predictor scores, but success is based either on the true scores of the predictor (i.e., selection accuracy) or the true scores on the criterion. Certain relative improvements in selection can be determined by comparing across simulation conditions that vary in their parameters.

The current project extended previous findings by examining the effect of measurement unreliability on classification accuracy, combined with its implications for productivity loss for the organization, where the alpha reliabilities of each predictor in the composite vary and are based on realistic conditions suggested by the organizational literature for what realistically might be found in job applicant data in a personnel selection setting: .81 for cognitive ability (Hattrup, O'Connell, & Labrador, 2005), .84 for structured interview (McDaniel, Whetzel, Schmidt, & Maurer, 1994), .78 for conscientiousness (Viswesvaran, & Ones, 2000), and .79 for biodata (Dean, 2004). Using alpha as indicator of reliability for structured interview and biodata is worth noting. One aspect in the literature about alpha is that it is a measure of the extent to which a general unidimensional construct is present among the items (Crano

& Brewer, 1973; Hattie, 1985). From this perspective, alpha is not an ideal indicator of reliability for structured interview and biodata because they are measurement methods that usually measure a number of different constructs, rather than measures of a specific unidimensional construct domain (constructs that they measure is dependent upon the types of items that are administered). However, a general unidimensional factor is a necessary but not a required property for high alpha, which is a function of internal consistency, or the degree of interrelatedness among the items (Cortina, 1993; Crano & Brewer, 1973). Thus, even if the items do not uniformly load onto a single factor, to the extent that there is close factor interrelatedness and low item-specific variance (uniqueness associated with the items), test is psychologically interpretable and internally consistent (Cronbach, 1951). Thus, alpha can be a useful index of reliability for structured interview and biodata.

Terminology. The current work requires some terminology for making selection decisions with less-than-perfectly reliable indicators of the work-relevant constructs that they purport to measure. Making correct and incorrect selection decisions due to measurement error variance in the predictor battery will be called *selection accuracy*. Falling under the umbrella of selection accuracy, there are two types of *selection successes*: Accepting and rejecting those whose observed scores on the predictor battery would have led to the same decisions on their corresponding true scores on the predictor battery (i.e., *true accepts* and *true rejects*, respectively). There are also two types of *selection errors*, where the selection decision based on the predictor battery of observed scores is the opposite of the decision that would have been made based on the predictor battery of true scores (if they could be known). Namely, one might reject applicants who should have been hired or hire applicants

who should have been rejected; these selection errors are called *false rejects* and *false accepts*, respectively (see Figure 2).

In reality, selection decisions can only be made based on the unreliable observed predictor scores. Applicants who are actually selected or rejected will be called *observed score accepts* and *observed score rejects*, respectively. On the other hand, applicants that organizations would have hired or rejected if true predictor scores of the applicants would have been available to them will be called *true score accepts* and *true score rejects*, respectively.

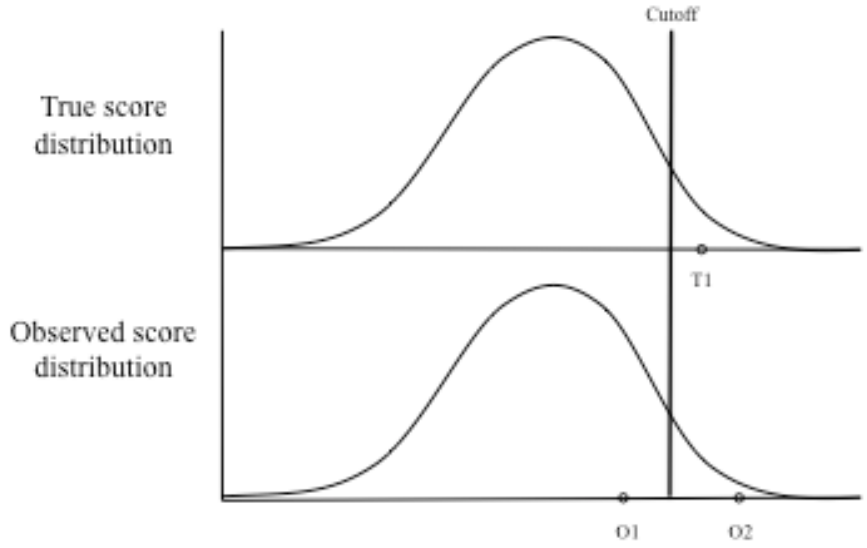

Figure 2. Illustration of selection accuracy based on score deviations that happen because of measurement error variance. Given applicant true score (T1), his/her corresponding observed score may be lower (O1) or higher (O2), with the magnitude depending on the degree of unreliability. It can be seen from this figure that applicant whose true score meets the cutoff set by the organization (T1) may be correctly selected (O2) or erroneously rejected (O1) based on the corresponding observed score.

Selection Parameters and Conditions

Combining multiple predictors. For selection purposes, organizations usually assess applicants using sets of predictors involving multiple measures or tests (Boudreau, Sturman, & Judge, 1994; Gatewood & Field, 2004) that purport to measure their standing on constructs that are intended to predict applicants' standing on the construct(s) that underlie job performance (Binning & Barrett, 1989). Regarding the method for combining predictor scores for use in top-down selection procedures, organizations can either compute a unit-weighted composite (e.g., standardize each variable and then average the *z*-scores together) or they apply a linear regression model, where each predictor variable receive differential weights based on predictor intercorrelations and criterion-related validity (Cascio, 1991; Gatewood & Field, 2004; Lord, 1962). Both cases are considered compensatory models, because higher scores on some predictors will compensate for lower scores in others (Potosky, Bobko, & Roth, 2005). For example, if scores on cognitive ability and conscientiousness measures are standardized, regression weighted, and combined for use in top-down selection, then higher cognitive ability scores will compensate for lower conscientiousness scores, and conversely, higher conscientiousness scores will compensate for lower ability scores.

Selection based on a compensatory model is appropriate in cases where different competencies are allowed to compensate when selecting candidates and when predicting a criterion outcome of interest (Kane & Case, 2004). For example, data might suggest that the same level of job performance might arise from two different profiles of predictor scores: higher levels of cognitive ability and lower levels of conscientiousness, or higher levels of conscientiousness and lower levels of cognitive ability.

The compensatory model is also appropriate in cases where there is a considerable construct overlap among the test components that are used in selection (Ben-David, 2000; Kane & Case, 2004). As the degree of construct overlap increases, variance in the scores of the test components become more interdependent. Thus, a high correlation among the test components might recommend aggregating test scores to form a unidimensional score from related constructs (Ben-David, 2000). The psychometric theory of internal consistency reliability would be consistent with this approach (see Nunnally, 1978)

More complex decision rules that move beyond this compensatory model are also possible, such as using multiple-hurdle model (Cascio, 1991; Gatewood & Field, 2004; Hills, 1971; Lord, 1962), where large numbers of applicants might be screened on tests that are less expensive to administer (e.g., ability and personality tests); then a smaller subset of screened applicants receive tests that are more expensive or time consuming to administer (e.g., interview, job simulation).

The rationale for implementing a multiple-hurdle model in selection is based on the grounds that a certain level of competency is required in all of the knowledge, skills, abilities, and other characteristics (KSAOs) that are measured by the predictor battery (Hills, 1971; Lord, 1962). In multiple-hurdle model, high level of competency in one area does not compensate for low level of competency in another for acceptable performance.

Regardless of whether predictor constructs are related, distinct or something in between, the compensatory model of selection tends to be more reliable than more complex decision rules (Kane & Case, 2004). To give one example, a single cutoff based on a composite of tests will tend to be more reliable than a multiple-hurdle system based on multiple cutoffs from multiple tests, because composite in the former

will tend to be longer and more reliable than each of the constituent tests in the latter (Hambleton & Slater, 1997). In a simulation study supporting this point, Douglas and Mislevy (2010) compared levels of selection accuracy, defined as the proportion of subjects who received the same classification based on true scores and observed scores, among several different selection models – compensatory, multiple hurdle, complementary (requires passing at least one of a number of tests), conjunctive-complementary (requires passing all tests, and at least one test in a different set of tests), and conjunctive-compensatory (requires passing all tests and attaining a prescribed total score) – with results indicating that the compensatory model generally provides the highest overall selection accuracy. Despite the lower reliability compared to compensatory model (Douglas & Mislevy, 2010; Hambleton & Slater, 1997), multiple-hurdle model is widely used in practice because it can save costs when making applicant assessments (De Corte, Lievens, & Sackett, 2006; Hills, 1971; Sackett & Roth, 1996).

The choice of selection process (e.g., compensatory vs. multiple hurdle) has important effect on who is selected from the applicant pool and the level of resulting performance scores (Chester, 2003). Thus, sensitivity to the nature of different decision rules for selection is important and should be implemented based on the specific purpose or needs of the organization. The current simulations focus on both compensatory and multiple-hurdle selection models to compare the use of a more reliable method (compensatory) vs. a more popular method (multiple-hurdle) on organizational benefits in terms of selection accuracy, predicted employee job performance, and ultimately any utility estimates derived from these outcomes.

Selection ratio. All other things being equal, the selection ratio affects the nature of selection decisions in terms of selection accuracy as well as the selection

errors that are inevitably made. More specifically, measurement error variance is negatively correlated with the observed score in the pool of those hired, given top-down selection and a specific selection ratio (Mendoza & Mumford, 1987). The lower the selection ratio, the higher the negative correlation; thus, more stringent selection generally leads to fewer false positives, but at the cost of increasing false negatives.

Methods for determining the selection decision point. The *Standards for Educational Psychology Testing* (1999), and the *Uniform Guidelines on Employee Selection Procedures* (1978) advise that when measurements are used for selection purposes, their critical scores should be set so as to reflect the level of knowledge or skills expected for acceptable performance within the occupation or profession. A cutoff score on the other hand, defines a specified point on the predictor score distribution below which candidates are rejected. Hopefully, the cutoff score an organization sets on its selection measures is higher than or at least equal to the critical score to ensure that non-qualified individuals are not selected. However, in selection situations where it is difficult to discern the absolute minimum level of knowledge or skills necessary for acceptable performance, cutoff score that the organization sets on its selection measures may or may not be aligned with the theoretical critical score. In such cases, even if the selection cutoff is at the point that is viewed as a critical score, selection errors (e.g., selection error due to measurement unreliability) are likely to occur and pose potentially hazardous consequences for the organization (SIOP, 2003).

In addition to making selection decisions based on specific critical or cutoff scores, another strategy is to select candidates in a top-down manner (SIOP, 2003). Given that there is a linear relationship between the predictor and the criterion, top-

down selection generally optimizes the level of predicted criterion performance (with the assumption that the predictor weighting is appropriate and there is an appropriate amount of variance in the predictor; Cascio, Outtz, Zedeck, & Goldstein, 1991). However, with this advantage of the rank ordering selection method comes the downside of increased potential for adverse impact when cognitive ability testing is involved, as has been demonstrated in a considerable amount of research (e.g., Bobko, Roth, & Potosky, 1999; Cascio et al., 1991; Hunter, Schmidt, & Rauschenberger, 1977; Roth, Switzer, Van Iddekinge, & Oh, 2011; Sackett & Roth, 1991; Sackett, Schmitt, Ellingson, & Kabin, 2001). Organizations may favor a particular alternative (maximum performance vs. diversity) depending on the goals or the purposes of the selection. For example, if the primary goal of the selection procedure is to minimize the potential for adverse impact, organizations might set a low boundary or reduce the regression weights on the cognitive ability score, and they might also set a more stringent boundary or increase the regression weights on the non-cognitive ability measure scores (e.g., conscientiousness), where racial group mean differences have been shown to be smaller compared to mean differences in cognitive ability measures (e.g., Barrick & Mount, 1991; Sackett et al., 2001; Sackett & Wilk, 1994). In doing so, racial/ethnic diversity in the selected population increases.

However, a strong caution against selecting applicants in this manner is if the validity of the selection battery is compromised. If cognitive ability is the most valid predictor of job performance, as is often found, then being less selective on this measure may reduce adverse impact but also will compromise validity (this is more likely to be the case to the extent that job performance is saturated with cognitive ability). For example, De Corte, Lievens, and Sackett (2007) demonstrated that non-cognitive ability measures had to be weighted substantially heavier relative to

cognitive ability measures before satisfying the 4/5ths rule. Consequently, expected standardized criterion performance of the selected applicants decreased from .78 when maximum importance was placed on job performance to .47 at the point where the selection procedure satisfied the 4/5ths rule. At this point, standardized predictor weight placed on cognitive ability decreased from .28 to .00, whereas the weight placed on conscientiousness increased from .12 to .79. To some extent, the tradeoff between adverse impact and validity is unavoidable (Ployhart & Holtz, 2008; Hoffman & Thornton, 1997). However, organizations should understand the practical consequences of using one selection method versus the other (e.g., tradeoff between performance and diversity; Sackett, De Corte, & Lievens, 2010), and they should be able to justify their rationale for implementing their selection method.

Current simulations will focus on two different types of top-down selection models: for the multiple-hurdle model, there is top-down selection at each hurdle; and for the compensatory model, top-down selection is based on the unit-weighted vs. regression-weighted composite scores. The assumption of top-down selection is that organizations are focusing on optimizing prediction of performance outcomes by selecting the best applicants out of a fixed pool. Simulations are based on the reasonable assumption that linear relationships exist between each of the predictors and the job performance criterion, though the strength of relationship (validity) varies by predictor, and predictors are themselves somewhat intercorrelated. Past findings have shown support for the linear relationship between both cognitive ability and conscientiousness with the job performance criterion (see Coward & Sackett, 1990, and Robie & Ryan, 1999, respectively). As methods, both structured interview and biodata have also shown consistent positive linear relationships with job performance criterion (McDaniel et al., 1994; Schmidt & Hunter, 1998). Thus, the simulation can

safely proceed from meta-analytic findings of correlations that indicate linear relationships between measures of the predictor and criterion constructs.

In the current simulations, selection ratio refers to the proportion of candidates from the population that is selected by the organization in a top-down manner. Because the selection purposes or needs vary both within and across organizations, the simulations likewise vary the selection ratio across a range of values.

Unit-weighted vs. Regression-weighted composite. In a compensatory selection model, organizations assign weights to performance predictors to form a single composite score. Weights can be derived by way of expert judgment, where subject matter experts (SMEs) are asked to indicate the relative importance of the predictors by assigning weights to them, and then a final set of weights (e.g., an average across SME judgments) is applied to each predictor in forming a composite. Weights can also be derived from empirical modeling, where linear regression weights are derived (either from the data on hand or from a different sample) and then applied to the predictors, with the goal of minimizing squared errors of prediction for a specific criterion. The SME and regression approaches are both legitimate approaches, but each has caveats that suggest they be used with care (e.g., not all SMEs generate the same weights; sensitivity of regression weights to capitalization on chance; collinearity between predictors).

Regarding the criterion-related validity of a composite score, the estimate of the multiple correlation in the population is greater as the pattern of the predictor ordinary least squares (OLS) regression weights approximate the corresponding population regression weights. This fact is true regardless of the differential pattern of criterion-related validity of each individual predictor variables with the criterion, intercorrelation among the predictor variables, and ratio of the weights assigned to the

predictor variables (Ghiselli, Campbell, & Zedeck, 1981), although all of these factors have critical effects on this approximation.

There is a long history of methodological studies examining the effect of different weighting schemes on the reliability and stability of a composite score (e.g., Meehl, 1954). As mentioned earlier, regression models tend to produce better prediction compared to expert judgment. Dawes and Corrigan (1974) replicated this finding. Interestingly however, they also found that a regression model outperformed human judgment even when random weights were applied to the predictors. Ree, Carretta, and Earles (1998) found similar results. In their study, they generated 11 different sets of randomly generated weights that were assigned to each test within the Armed Services Vocational Aptitude Battery (ASVAB; Ree & Carretta, 1994). The results showed that the correlations among the rank-order of the composite scores from differentially weighted predictor scores ranged from .97 to 1.00. Their study provided compelling evidence that reliability of the rank-order of the composite scores is generally robust. In fact, Wilks (1938) provided a theorem that mathematically showed the conditions under which differentially weighting the predictors has little influence on the rank-order of the composite scores. This is true if the regression weights are produced on predictors with even a moderate level of positive correlation, and the relative variability of the weights is not great; this is especially true as the number of predictors increases under these conditions. Results found in the literature generally have confirmed Wilks' (1938) theorem (e.g., Allen & Yen, 1979; Guilliksen, 1950).

Similarly, Wainer (1976) provided a mathematical proof that unit-weighting the predictor scores to form the composite score also had minimal effect on the accuracy of regression models. He also pointed to a number of benefits of using unit-

weights, including ease of estimation, and insensitivity of outliers or non-normality of the distribution. The current project focuses on comparing selection models that use unit-weighted and regression-weighted composite scores that vary in the reliability of those predictors.

Dimensionality of the job performance criterion. Van Iddekinge and Ployhart (2008) pointed that even though researchers find the idea of multidimensional criteria to be theoretically appealing, it has not been justified empirically. This is because the performance ratings criteria are highly intercorrelated. In fact, factor analytic studies that analyzed interrelationships among the criterion variables generally have shown evidence for a dominant single factor even after correcting for halo error, rater effects, random response error, and transient error (Viswesvaran, Schmidt, & Ones, 2005). Thus, many organizations and researchers use overall job performance ratings in practice (Van Iddekinge & Ployhart, 2008), including the meta-analytic findings that are used to generate the simulation data in the current thesis.

Selection Utility

Current simulations contribute to the past work on utility analysis in several ways. In I/O psychology, the utility of a selection battery is often associated with the dollar metric (e.g., Boudreau, 1983; Schmidt, Mack, & Hunter, 1984). Recall, however, that selection utility is a general term applied to mean the degree to which the selection procedure improves the quality of selection in comparison to what would have occurred had it not been used (SIOP, 2003). In the current thesis, selection utility of perfectly reliable predictor measures over that of unreliable predictor measures is illustrated in terms of a) relative proportion of the selection success and selection error in the predictor, b) mean difference in predictor scores (true vs.

observed predictor scores) between the selection decision groups, and c) their criterion performance (see Table 1 for summary of the group mean comparisons that are made).

The mean differences in the predictor and criterion scores are illustrated through standardized group mean difference (mean group difference between selection decision groups over *total standard deviation* of the population – 1.0), and Cohen's *U* statistics. Cohen's *U* statistics measure the percentage of overlap (or non-overlap) between two populations (assuming normality and equal variability) in terms of *d* (Cohen, 1988). Specifically, U_3 is used to report the selection utility in terms of performance difference between the comparison groups. U_3 measures the percentage of population distribution of the lower scoring group that the mean of the higher scoring group exceeds. For example, if the mean of Group B is two standard deviations above the mean of Group A (i.e., $d = 2.0$), upper half of the distribution of Group B exceeds 97.7% of the Group A distribution, so $U_3 = 97.7\%$.

Table 1
Standardized Mean Comparisons Made

Mean Comparisons in the Predictor
Observed predictor scores for the *observed score accepts* vs. Observed predictor scores for the *observed score rejects*
True predictor scores for the *observed score accepts* vs. True predictor scores for the *observed score rejects*
Mean Comparisons in the Criterion
True criterion scores for the *true score accepts* vs. True criterion scores for the *observed score accepts*
True criterion scores for the *observed score accepts* vs. True criterion scores for the *observed score rejects*
True criterion scores for the *true score accepts* vs. True criterion scores for the *true score rejects*

Notes. *Observed score accepts* = applicants who were selected based on observed predictor scores; *Observed score rejects* = applicants who were rejected based on observed predictor scores; *True score accepts* = applicants who should have been selected had selection been based on true predictor scores; *True score rejects* = applicants who should have been rejected had selection been based on true predictor scores.

Selection accuracy. The current thesis presents results on *selection accuracy,* or how much specific influence measurement error variance (unreliability) has on selection utility in terms of the proportion of applicants who would receive the same selection decision (accept or reject) based on true scores and observed scores on a given set of predictors. Because of measurement unreliability, the rank-order of the applicants on the observed predictor scores will not perfectly correspond to the rank-order on their respective true predictor scores. Consequently, any top-down selection on the observed predictor score will not capture all of the best applicants as reflected by their true scores (i.e., random errors of measurement necessarily result in errors of selection). Consider a hypothetical situation where employees are selected based on the observed scores of a test. To the extent that the cutoff point on the predictor is higher (i.e., the selection ratio is lower), the proportion of measurement error variance in the selected group will be reduced (Mendoza & Mumford, 1987). Therefore, when organizations are more selective, they are saying that a higher true score is important to them, and as a result, there is a lower likelihood of observing a score deviation that results in false accepts *and* false rejects; conversely, when the selection ratio is higher and organizations are less selective, there is a greater likelihood of both types of errors (see Figures 3 and 4). This illustrates the regression-to-the-mean effect in the relationship between observed scores and true scores, as it applies to the selection setting.

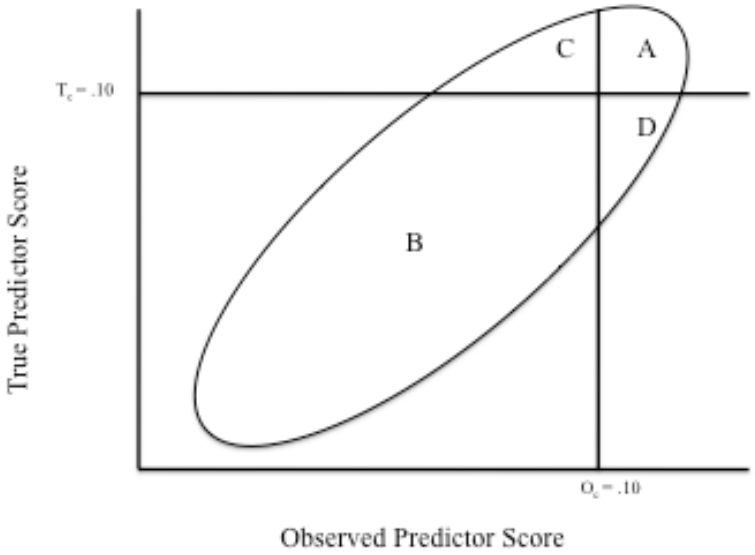

Figure 3. Selection accuracy when organizations are highly selective (SR = .10). T_c and O_c are cut points on true and predictor scores, respectively. A = true accept region; B = true reject region; C = false reject region; D = false accept region. The more the SR deviates from .50, the higher the proportion of classification accuracy (A + B) is relative to classification errors (C + D).

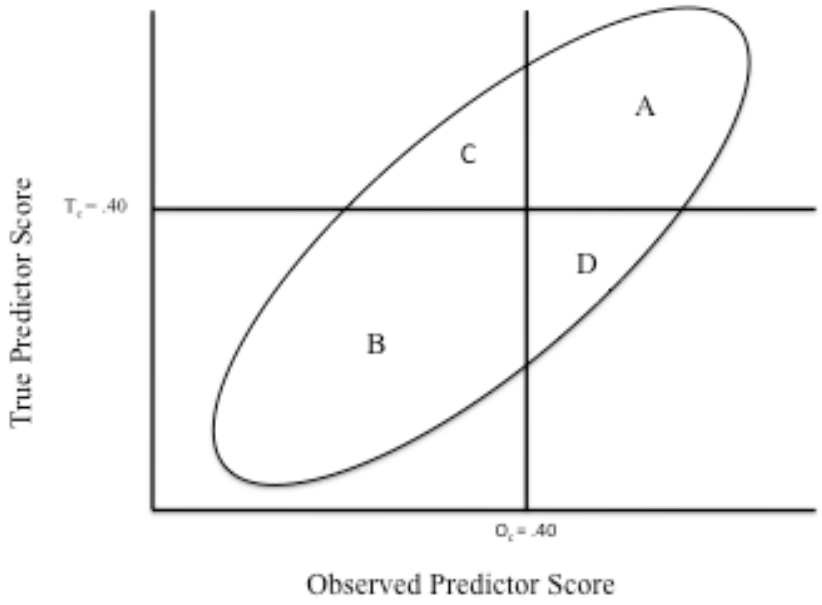

Figure 4. Selection accuracy when organizations are less selective (SR = .40). T_c and O_c are cut points for true and predictor scores, respectively. A = true accept region; B = true reject region; C = false reject region; D = false accept region. When SR is high, relative proportion of classification accuracy (A + B) is lower compared to classification error (C + D).

For the multiple-hurdle selection, *true accepts* were defined as applicants who were selected based on their true predictor scores and observed predictor scores on all of the selection hurdles. On the other hand, *false accepts* were defined as applicants who were selected based on their observed predictor scores, but would not have been selected on at least one of the selection hurdles based on their true predictor scores. I investigated the rate of selection success (*true accepts*) and selection errors (*false accepts*) only among the selected applicants because of the difficulty involved in distinguishing the difference between *false accepts* and *false rejects*. Considering the fact that erroneously hired applicants are in an immediate position to potentially

slow or even hurt the organizational performance, organizations may be more likely to be interested in the rate at which selection errors occur among the selected applicants. The proportion of selection success (or selection error) illustrates the rate of selection accuracy that a selection battery with a given measurement reliability can produce.

Selection outcomes are to be contrasted with the more commonly identified *prediction accuracy* that matches applicants' observed predictor scores against their respective criterion scores (e.g., if fallible predictors are used operationally to predict true levels of performance; Binning & Barrett, 1989). Under the umbrella of prediction accuracy, *prediction successes* happen when hired applicants meet the expected standard of performance or applicants who are not hired would not have reached this standard; these successes are called *true positives* and *true negatives*, respectively. Conversely, *prediction errors* happen when hired applicants do not meet standards, or rejected applicants would have met the standard; these errors are called *false positives* and *false negatives*, respectively. These prediction accuracy rates are not computed in the current simulations because I am not dichotomizing performance criterion scores (success vs. failure). This dichotomizing might be useful for thinking about percentage of selection success from the selection process (as is done in the Taylor-Russell model), but organizations often do not set these specific standards or those standards are based on arbitrary managerial consensus (Cascio, 1980) that may not be generalizable to different job contexts.

Mean comparisons on the predictor. Because of measurement unreliability, the correlation between the observed predictor scores and the true predictor scores is less than 1.00. Therefore, when selection is based on the observed scores of the predictors, there will be selection errors, where mean true scores of those

falsely selected will be lower than their mean observed scores. Similarly, the mean true scores of those falsely rejected will be higher than their mean observed scores. The purpose of this comparison is to illustrate that utility in terms of mean predicted performance that is thought to be gained through using the selection tool should be more conservative than is suggested by the difference in the mean observed predictor scores between the selected group and the rejected group.

To illustrate the effect of measurement unreliability on mean predictor scores, I compared the mean observed and true predictor scores between the *observed score accepts* (*true accepts* and *false accepts*) and the *observed score rejects* (*true rejects* and *false rejects*). Same comparisons were made in the multiple hurdles model condition, where the selected group comprised of applicants who passed all selection hurdles based on their observed predictor scores, and the rejected group comprised of applicants who did not (mean comparisons between selection decision groups were based on the applicants' composite predictor scores rather than their predictor scores on a specific selection hurdle).

Mean comparisons on the criterion. Not only are predictor scores imperfect indicators of their respective true scores; predictor batteries should never be expected to correlate perfectly with the organizational outcome of interest. Therefore, in addition to measurement errors in the predictor, further selection errors occur when the applicants' observed or true predictor scores are used to predict respective true criterion score. In turn, *selection errors* have implications for subsequent criterion performance, as reflected by average true scores on the criterion for those who were selected. Rather than modeling the relative rates at which selection errors occur in the criterion, the focus is on how measurement unreliability affects the outcomes of selection in terms of mean difference in the true criterion performance between when

selection is based on observed predictor scores versus when selection is based on true predictor scores.

Latent constructs, such as job performance, are impossible to measure directly. Thus, we make inferences about test-takers' true standing on the focal ability based on the observed scores from some form of indirect measure of the construct (Binning & Barrett, 1989). Although organizational decisions about HR practices (e.g., promotion, termination) are based on observed performance scores, current thesis is interested in predicting the theoretical standing on the job performance construct.

For both compensatory and multiple-hurdle model selections, I first compared the difference in mean criterion performance between the *observed score selects* and the *true score selects*. Any difference in this comparison illustrates the gains in utility in terms of mean criterion performance that organizations could achieve from selections based on perfectly reliable predictor measures in comparison to selections based on unreliable predictor measures. I then compared the difference in mean criterion performance between the *observed score accepts* and the *observed score rejects* versus the difference between the *true score accepts* and the *true score rejects*. The magnitude to which the difference in criterion performance between the selection decision groups is greater for the *true score accepts – true score rejects* comparisons than to *observed score accepts – observed score rejects* comparisons illustrates the improvement in distinction between the selection decision groups that could be made from selection based on perfectly reliable predictor measures in comparison to selections based on unreliable predictor measures.

The overall objective of this project is to provide a sample illustration of the effect that the aforementioned effects of the predictor reliability; criterion reliability

and validity; selection ratio; and intercorrelations and weighting regression have on the accuracy of selection decisions, and outcomes of selection in terms of the mean difference in the predictor score and criterion performance between the selection decision groups. Note that the values of the parameter estimates (e.g., validity coefficients, reliability estimates) or the assumptions made in the current simulations (e.g., conceptualization of the variables) may differ considerably depending on the specific context or theory. Rather viewing the assumptions underlying the characteristics of the selection situations and the resulting outcomes as the definitive summary, they should be viewed as an illustration of the general principles of the effect that measurement unreliability could have across a range of selection situations for a range of different jobs.

Method

The current simulations are based on a population correlation matrix from Roth et al. (2011), containing predictor and criterion variables relevant to personnel selection: cognitive ability, structured interview, conscientiousness, and biodata as predictors, and an overall job performance criterion. (This correlation matrix updates a similar one generated from the meta-analysis by Bobko et al. (1999). Critically, Roth et al. improved on Bobko et al. (1999) by including more recent meta-analytic findings, making more refined distinctions (e.g., validities for high vs. low-complexity jobs) and attempting to correct correlations for range restriction to reflect relationships in the job applicant pool (vs. the more selective incumbent pool that tends to reflect attenuated correlations). For validity estimates, they also substituted the observed validity coefficients with operational validities (validities corrected for both range restriction and criterion unreliability); for variable intercorrelations, they substituted the observed correlations with meta-analytic studies between each pair of

the predictor variables, correcting them for range restriction. Note that two predictor intercorrelations (structured interview – biodata and conscientiousness – biodata) could not be corrected due to lack of appropriate information on range restriction.

Based on this matrix, Roth et al. (2011) then generated predictor and job performance criterion data based on both the uncorrected and corrected matrices and for each matrix conducted a multiple regression analysis by regressing job performance on the predictor variables. Results showed that, as would be expected, using the corrected input matrix greatly increased the multiple correlation coefficient (R = .75 for medium-complexity jobs) compared to the multiple correlation coefficient based on the uncorrected input matrix (R = .48). Perhaps somewhat less expected was the way in which the pattern of beta weights changed between corrected and uncorrected matrices. Based on the corrected input matrix, the beta weight for cognitive ability increased from .20 (uncorrected) to .40, whereas the beta weight for biodata decreased from .14 (uncorrected) to .00 (i.e., biodata showed no incremental validity in the context of the full regression model). Our results will differ from Roth et al. because we are also considering predictors corrected for measurement error variance, and we are examining a different level of unreliability for the performance criterion.

Input correlation matrix. Table 2 shows the correlation matrix used in the current simulation. Values below the main diagonal are the observed correlations from Roth et al. (2011) that are corrected only for range restriction. All of these values are attenuated by unreliability of measurement, except the operational validities from Roth et al., which are corrected for criterion reliability. The current simulation requires all of the observed correlations in the applicant pool; therefore the operational validities were attenuated based on reported criterion reliability values

that Roth et al. used with respect to each criterion. Although these values were not reported, I was able to obtain them by referring back to the meta-analyses that Roth et al. cited. The criterion reliability estimates associated with each predictor were: .60 (Hunter, 1986) and .52 (Salgado, Anderson, Moscoso, Bertuna, & de Fruyt, 2003a; Salgado et al., 2003b) for cognitive ability, .60 for structured interview (Potosky et al., 2005), .59 for conscientiousness (Hurtz & Donovan, 2000), and .64 for biodata (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). The average of these values for criterion unreliability across all predictors were used for correction (.59).

Next, values above the main diagonal in Table 2 are true correlations that disattenuate the observed validities (below the diagonal) by reliability estimates (alphas) in both the predictor and the criterion. Thus, these correlations are corrected for the effects of both range restriction and the reliability of each variable involved in the correlation (i.e., predictor-predictor or predictor-criterion correlation).

Criterion-related validity for cognitive ability. Recognizing that job complexity has important theoretical implications for understanding criterion-related validity estimates for cognitive ability in employment settings (Hunter 1986), Roth et al. (2011) reported two separate validity estimates: one for medium-complexity jobs, and one for low-complexity jobs. They obtained these values by averaging validities by level of job complexity across the studies by Hunter (1986), and Salgado et al. (2003a; 2003b). In the current simulation that bases itself on overall estimates of validity, I decided to average the meta-analytic validities Roth et al. report across all three levels of job complexity (high-, medium-, and low-complexity). Although job complexity is an important moderator to consider, and although it is generally important for simulation studies to use multiple values to be able to illustrate the range of potential outcomes that might arise across contexts (Schmitt, Rogers, Chan,

Sheppard, & Jennings, 1997), general estimate of the validity for cognitive ability is sufficient to illustrate the purpose of the current study. In addition, all other correlations in Roth et al. do not take job complexity into account.

Table 2
Observed and Corrected Zero-Order Correlations between Four Predictors and Job Performance

Measure	1.	2.	3.	4.	5.
1. Cognitive ability	(.81)[a]	.38	.04	.46	.57
2. Structured interview	.31	(.84)[b]	.16	.20	.53
3. Conscientiousness	.03	.13	(.78)[c]	.65	.25
4. Biodata	.37	.16	.51	(.79)[d]	.37
5. Job performance	.40	.37	.17	.26	(.59)

Notes. Alpha reliabilities are shown in parentheses along the main diagonal. Observed correlations from Roth, Switzer, Van Iddekinge, and Oh (2011) are below the main diagonal. These correlations reflect job incumbent correlations that were corrected for range restriction, with the exception of the biodata-structured interview and biodata-conscientiousness correlations, where authors could not find appropriately corrected correlations. True correlations are above the main diagonal; these are also corrected for measurement error variance in the predictor and criterion (one can attenuate these by predictor reliability to obtain operational validities). Alpha reliabilities come from the following sources: [a] for cognitive ability, see Hattrup, O'Connell, and Labrador (2005); [b] for the structured interview, see McDaniel, Whetzel, Schmidt, and Mauer (1994); [c] for conscientiousness, see Viswesvaran, and Ones (2000); [d] for biodata, see Dean (2004)

Thus, to illustrate some general principles, the current thesis focuses on a relatively specific and limited case for illustration, based on a reasonable and current set of correlations that are derived from the Roth et al. (2011) meta-analysis of employment research: Table 2 shows the correlations between four predictors and one criterion that are used to generate data with a multivariate normal distribution. I generated sample realizations of true scores as well as their corresponding observed scores, given the reliability coefficients and intercorrelations for these five variables (see Kaiser & Dickman, 1962, for the singular value decomposition method employed). Predictor composite scores and criterion scores were standardized within the true score data and observed score data, to ensure comparability.

Mathematically, if the observed score is X_O, the true score is X_T, and the

population reliability is r_{xx}, then the formula for generating observed scores in the sample is:

$$X_o = \sqrt{r_{xx}}(X_T - \bar{X}_T) + \dot{s}_x\sqrt{1 - r_{xx}}$$

, where $\bar{X}_T$ is the mean of the true scores in the sample, and $\dot{s}_x$ is a score from a random normal distribution with a mean of zero and a standard deviation of one. The equation shows how this standard deviation gets adjusted by the standard error of measurement; the higher the reliability, the lower the standard deviation of the error score.

The simulation then performs top-down selection on the observed predictor scores three times, once based on the unit-weighted predictor composite, once based on the regression-based predictor composite, and once for multiple-hurdle selection (see Table 3 for a list of assumptions made in the current simulation). Selection accuracy is first examined, comparing selection based on applicants' true predictor scores vs. their observed predictor scores and identifying applicants who were correctly selected or correctly rejected based on their observed predictor scores (*true accepts* and *true rejects*, respectively), and also identifying applicants who were incorrectly selected or incorrectly rejected on the observed predictor score (*false accepts* and *false rejects*, respectively, where the true score indicates the opposite selection decision should have been made). These comparisons were made for compensatory selection models that vary in the selection ratios (SR = .10, .20, .40), as well as for the multiple-hurdles model (to be described shortly). Mean levels of true and observed predictor scores, as well as true criterion scores, were then calculated for each type of selection decision (i.e., *selected* vs. *rejected*). I also varied the number of applicants across three levels (N = 250, 500, 1000). Although the average results across 1,000 replications should essentially be the same regardless of the

sample size, varying sample sizes for each condition allows one to understand how much one can expect results within a single selection setting (replication) to vary from the expected value. This is obviously important, because an organization wants to know what value to expect in *their* situation, and their situation may not yield the average result when there is a lot of associated variability due to a smaller sample size. As outlined in Table 3, I examined 108 different combinations of selection ratio, number of applicants, selection decision rules (unit-weighted compensatory, regression-weighted compensatory, multiple hurdles), predictor reliability (true and observed), and criterion reliability (true and observed) in a 3 × 3 × 3 × 2 × 2 factorial design.

Table 3
Characteristics and Parameters of Selection Simulations

Predictors
Cognitive ability, conscientiousness, biodata, structured interview
Criterion
Overall job performance
Selection Method
Compensatory vs. multiple hurdles
Parameters
Compensatory Selection (two methods)
Regression-weighted
Unit-weighted
Multiple-hurdle Selection(one method)
Hurdle 1: SR = .70 based on cognitive ability scores
Hurdle 2: SR = .50 based on conscientiousness + biodata (unit-weighted) scores
Hurdle 3: SR = .20 based on structured interview scores
Predictor reliability:
True scores (perfect reliability)
Realistic reliability: .81 for cognitive ability; .84 for structured interview; .78 for conscientiousness; .79 for biodata
Criterion reliability
True scores (perfect reliability)
Realistic reliability: .59
Selection ratios: .10, .20, .40
Number of applicants: 250, 500, 1000

Note. For each condition, simulations are replicated 1,000 times, then the mean and standard deviation of the results across replications are reported.

Multiple-hurdle process. For practical reasons, organizations often screen a large number of initial applicants with measures that are relatively quick and inexpensive to administer, and then apply more time-intensive and costly measures on a more limited sample (Schmidt & Hunter, 1998). Thus, cognitive ability related measures are often used as the initial screening tool given their validity and low costs (Roth, Bobko, Switzer, & Dean, 2001). On the other hand, structured interviews are often used in the later part of a selection procedure because of the higher requirements of time, cost, and effort. Consequently, many of the organizations in the public sector (e.g., Dipboye, Gaugler, Hayes, & Parker, 1999; Distefano & Pryer, 1987), and the private sector (e.g., Kaiser, Adorno, Williams, & Binning, 1996; Roth & Campion, 1992; Sackett & Wilk, 1994) use cognitive ability measure as the initial screening process before interviews.

The multiple-hurdle simulation is aligned with these notions by assuming that 70% of the applicants are selected on the initial cognitive ability measure; then 50% of the applicants who passed the cognitive ability test are selected on the combined conscientiousness and biodata scores in the second hurdle, and finally, 20% of the applicants who passed the second hurdle are selected on the structured interview in the final hurdle. Overall, 7% of the original applicant sample is selected from the multiple-hurdle selection process.

There are certainly a number of other potential selection scenarios that are justifiable for conceptual, practical, and legal reasons. The assumptions in the current simulations certainly do not reflect the wider range of parameters or the additional factors that could be incorporated (some of which are mentioned in conclusion). Instead, the purpose of the current simulations is to illustrate the general principle about the practical influence that measurement artifacts have on selection under a

realistic set of conditions implied from meta-analysis of variables relevant to personnel selection.

Procedures

Simulations were programmed using R Code (R Development Core Team, 2009). Population regression weights are based on my version of the Roth et al. (2011) correlation matrix, as previously described, in order to calculate the predictor beta-weights for true scores and observed scores, respectively (i.e., **B**=inv(**Rxx**)**Rxy**, where **Rxx** and **Rxy** are the $p \times p$ and $p \times 1$ partitions of either the corrected or uncorrected correlation matrix for p predictors and one criterion). Within each combination of the simulation parameters, I replicated the aforementioned selection process 1,000 times to calculate (a) the mean rate of selection for each type of selection decision on the true predictor score (*true accept*, *true reject*, *false accept*, *false reject*), (b) the mean predictor score, and (c) the mean true criterion score for the different selection decision groups. Certainly these mean values could be obtained from one large sample rather than 1,000 replicates. However, replicates are useful because of the expected variability in results across samples of the same size. An organization will have a single sample, and as such, it may not achieve results that are the same as a mean value; their results will deviate in ways consistent with the variability to be expected in that situation. Therefore, for each selection scenario studied in the simulation, the standard deviations of these results across replications index the variability to be expected for a given sample size and selection condition.

Initial Findings

Under the compensatory selection method, I examined the mean rate of selection accuracy (*true accepts* and *true rejects*) and selection errors (*false accepts* and *false rejects*), and the mean predictor scores (*true* vs. *observed*) within each type

of selection decision. I examined 12 different combinations of three factors: the selection ratio (SR = .10, .20, .40), predictor reliability (unreliable vs. perfectly reliable) and composite predictor weighting (regression-weighted vs. unit-weighted) in a 3 × 2 × 2 factorial design. I assumed a selection situation where the number of applicants is 250 and the criterion reliability is .65. Note that Roth et al.'s (2011) findings were not yet available when these initial results were generated. Therefore, the parameter estimates in the input true correlation matrix that was used to generate the true score data was only corrected for predictor unreliability listed above.

Selection on True vs. Observed Predictor Scores

Results compare applicants who would have been selected based on their true predictor score composite versus their observed predictor composite score (see Table 4). As was expected, selection errors in the predictor (i.e., false accepts and false rejects) increased as the selection ratio increased (i.e., less selectivity). However, it is worth noting that the increase in the rate of errors was small (see Table 5).

Table 4
Standardized Mean Predictor Composite Scores

Unit-Weighted Composite

	True scores		
	Selected?	Yes	No
Observed scores	Yes	SR=.10: 1.84 (.08) SR=.20: 1.48 (.05) SR=.40: 1.04 (.02)	SR=.10: 1.04 (.10) SR=.20: .61 (.07) SR=.40: .03 (.07)
	No	SR=.10: 1.47 (.09) SR=.20: 1.04 (.07) SR=.40: .47 (.06)	SR=.10: -.23 (.01) SR=.20: -.40 (.01) SR=.40: -.71 (.02)

Regression-Weighted Composite

	True scores		
	Selected?	Yes	No
Observed scores	Yes	SR=.10: 1.84 (.08) SR=.20: 1.48 (.05) SR=.40: 1.05 (.02)	SR=.10: 1.03 (.10) SR=.20: .60 (.08) SR=.40: .02 (.07)
	No	SR=.10: 1.48 (.09) SR=.20: 1.04 (.07) SR=.40: .47 (.06)	SR=.10: -.23 (.01) SR=.20: -.40 (.01) SR=.40: -.71 (.02)

Note. $N = 250$.

Table 5
Percent of Selection Success and Selection Errors

Unit-Weighted Composite

	True scores		
	Selected?	Yes	No
Observed scores	Yes	SR=.10: .08	SR=.10: .02
		SR=.20: .16	SR=.20: .04
		SR=.40: .34	SR=.40: .06
	No	SR=.10: .02	SR=.10: .88
		SR=.20: .04	SR=.20: .76
		SR=.40: .06	SR=.40: .54

Regression-Weighted Composite

	True scores		
	Selected?	Yes	No
Observed scores	Yes	SR=.10: .07	SR=.10: .03
		SR=.20: .16	SR=.20: .04
		SR=.40: .34	SR=.40: .06
	No	SR=.10: .03	SR=.10: .87
		SR=.20: .04	SR=.20: .76
		SR=.40: .06	SR=.40: .54

Note. $N = 250$. All standard deviations for the proportions across 1,000 replications are < .01.

Table 6 shows the true and observed standardized mean predictor scores for all selection decisions. The results across selection ratios generally showed that there was a distinct difference between the mean true predictor scores and mean observed predictor scores between those selected and rejected (based on observed predictor scores). When selection is based on the unreliable predictor scores, then the corresponding mean true predictor scores are attenuated (similar to the attenuation of standardized mean differences that are corrected for in meta-analysis); thus, the difference between the mean true predictor score on the composite between those selected and rejected was smaller. Table 6 shows that for *true accepts*, their standardized mean scores on the unreliable predictor scores and their corresponding true scores are the same. For *false accepts* however, their mean true scores are attenuated from their observed scores, therefore lowering the mean true score of the

selected group. Similarly, for *true rejects*, their standardized mean scores on the unreliable predictor scores and their corresponding true scores are the same. For *false rejects* however, their mean true scores are higher than their mean observed scores, therefore increasing the mean true score of the rejected group. These results directly reflect the influence that predictor unreliability has on selection errors. For *true accepts* and *true rejects*, their mean level of predictor scores remained the same regardless of whether selection is done on true or observed predictor scores. However, true score of the *false accepts* was lower than their observed score, and true score of the *false rejects* was reciprocally higher than their observed score. These score differences reflect what the organization could have gained through selection on perfectly reliable predictors.

Although this basic result is based on one focused selection scenario, it can provide a wealth of information about mean differences, selection errors, and selection decisions that will be provided in greater detail within the full analysis that is to follow.

Table 6
Standardized Mean Composite Scores for each Selection Condition

Mean Observed Scores				
	True Accepts	*False Accepts*	*True Rejects*	*False Rejects*
SR = .10	1.84	1.48	-0.23	1.03
SR = .20	1.49	1.05	-0.40	0.60
SR = .40	1.04	0.47	-0.71	0.03
Mean True Scores				
	True Accepts	*False Accepts*	*True Rejects*	*False Rejects*
SR = .10	1.84	1.03	-0.23	1.47
SR = .20	1.49	0.60	-0.40	1.05
SR = .40	1.04	0.03	-0.71	0.47

Unit-Weighted vs. Regression-Weighted Composite

The percent of selection success and selection error, and their mean predictor composite scores were almost identical for each condition, whether selection was

based on a unit-weighted or regression-weighted composite score. Consistent with what was expected from regression-to-the-mean effect, the rate of *false accepts* decreased as the selection ratio decreased.

Even this single scenario provides an illustration of how predictor reliability affects the practical outcome of selection accuracy. To the extent that a predictor composite is able to accurately capture the test-taker's true standing on the constructs being measured, both organizations and applicants will benefit by avoiding erroneous selection decisions. Future simulations that extend this work will generalize the present selection scenario those that vary the number of predictors as well as the pattern and level of predictor reliabilities, intercorrelations, and criterion-related validities.

Discussion of the Initial Findings

Past simulations have illustrated the important effects that psychometric properties of measures (e.g., measurement reliability, criterion-related validity) and organizational selection practices (e.g., complex decision rules, selection ratio) together have on the practical outcomes of personnel selection. However, these effects often get lost in translation when attempting to communicate how they affect personnel selection strategies in practical terms. Although several studies have examined the effect of measurement reliability on selection accuracy, they have been based on single tests and not situations typical in personnel selection, where a selection decision is based on various combinations of selection measures. Douglas and Mislevy (2010) acknowledged this issue and conducted a simulation study to examine the effect of measurement reliability on selection accuracy based on multiple tests. However, they assumed a hypothetical situation where the reliability coefficient for the predictors was uniformly high (r_{xx} = .90) and above a level that is more typical.

The purpose of the current study was to examine how the types and levels of measurement and situational factors relevant to personnel selection combine in the aggregate to influence selection decisions.

As expected by the regression-to-the-mean effect when selecting observed predictor composite scores in a top-down manner, and range restriction in the error variance, a lower selection ratio was associated with lower rate of *false accepts* and *false rejects* with respect to true predictor composite scores. This is evident in that the difference in the standardized mean composite score between the selected group (*true accepts* and *false rejects*), and the rejected group (*true rejects* and *false accepts*) continue to increase as the selection ratio decreased for all conditions considered in the study (mean difference of 1.26 when the selection ratio = .10; mean difference of 1.09 when the selection ratio = .40). In other words, there was a more distinctive difference between the selected group and rejected group with lower selection ratio because there were fewer selection errors. In addition, I illustrated the effect that relatively high measurement reliability (composite alpha = .90) has on selection accuracy and mean composite score of the selected and unselected applicants. The results showed that organizations may erroneously hire up to 6% of *false accepts* and also mistakenly turn down 6% of *false rejects* when the selection ratio is 40% for both unit-weighted and regression-weighted composite. This contrasts with 2% error rate for *false accepts* and *false rejects* when the selection ratio is 10%. The results showed that error rates decreased as the selection ratio decreased. Another interesting finding was that the pattern of results was almost identical in both unit-weighted and regression-weighted composite scores. This finding is aligned with the classic notion (Gulliksen, 1950) that different weighting systems have minimal impact on composite scores when many predictors are used and the predictors are highly correlated. This

study used four predictors, and the intercorrelation among the predictors ranged from $r = .00$ to $r = .65$ (corrected for measurement reliability). This evidence may indicate which combination of the number of predictors and predictor intercorrelations will distinguish the effectiveness or indifference of various weighting schemes on the composite score reliability.

Research Hypotheses

In addition to the expected increase in selection errors with lower measurement reliability, logical hypotheses regarding the influence of the selection parameters previously discussed follows:

H1a: Selection errors (false accepts and false rejects) will increase as the selection ratio increases because with greater proportion of applicants being selected, there is higher likelihood of misclassification on selection.

Also, a longer composite for a single cutoff based on a composite of tests will tend to be more reliable than each of the constituent tests in a multiple-hurdle selection based on multiple cutoffs. Thus,

H1b: Selection based on a compensatory model will be more reliable compared to selection based on a multiple-hurdle model.

As a consequence of selection errors due to predictor unreliability, expected gains in predicted performance through the use of the selection battery will be smaller compared to actual gains in predicted performance implied from predictor scores. Thus,

H2: The difference in true mean predictor score between the selected group and the rejected group will be smaller compared to the difference in observed mean predictor score between the selected group and the rejected group.

Predictor batteries are not perfectly correlated with job performance. Thus, selection errors in the predictor from measurement unreliability will translate into further selection errors in the criterion. Consequently, it is expected that gains in true mean criterion performance through selection on the observed predictor scores will be smaller than the gain in true mean criterion performance through selection on the perfectly reliable predictor scores. Thus,

H3a: The true mean criterion performance for the observed score accepts (applicants selected based on observed predictor scores) will be lower than the true mean criterion performance for the true score selects (applicants selected based on true predictor scores).

H3b: Similarly, there will be greater difference in mean true criterion performance between the true score accepts and true score rejects than difference in mean true criterion performance between the observed score accepts and observed score rejects.

Regarding regression weighting, based on previous findings, Ghiselli et al. (1981) argued that unless there (a) is a wide variation in the weights applied to each item, (b) low intercorrelations among the items, and (c) small number of items, differential weighting is unlikely to have significant effect on the validity or reliability of the measure. Therefore,

H4: Selection accuracy, mean predictor score, and mean criterion performance will be similar either when regression weights are applied or when unit weights are applied.

Results

Selection on True vs. Observed Predictor Scores

Compensatory model. Tables 7 and 8 show the simulation results for selection accuracy in the predictor for the two compensatory model selection conditions (regression- and unit-weight conditions). In line with the initial findings, selection success rates (true accept rates and true reject rates) decreased and selection error rates (false accept rates and false reject rates) increased as the selection ratio increased. Not only was this general trend consistent regardless of whether the predictor scores were unit-weight or regression-weighted, the value of the rates in each of the corresponding conditions were nearly equivalent. Although not central to the purposes of the current thesis, results also showed that the number of applicants did not have much influence on the mean selection accuracy rates.

Table 7
Percent of Selection Successes and Selection Errors

Regression-Weighted Composite					
SR = .10	*N*	*TA%*	*TR%*	*FA%*	*FR%*
	250	6.3 (.7)	86.3 (.7)	3.7 (.7)	3.7 (.7)
	500	6.3 (.5)	86.3 (.5)	3.7 (.5)	3.7 (.5)
	1000	6.3 (.4)	86.3 (.4)	3.7 (.4)	3.7 (.4)
Mean		6.3 (.6)	86.3 (.6)	3.7 (.6)	3.7 (.6)
SR = .20	250	13.9 (1.0)	73.9 (1.0)	6.1 (1.0)	6.1 (1.0)
	500	14.0 (.7)	74.0 (.7)	6.0 (.7)	6.0 (.7)
	1000	14.0 (.5)	74.0 (.5)	6.0 (.5)	6.0 (.5)
Mean		14.0 (.8)	74.0 (.8)	6.0 (.8)	6.0 (.8)
SR = .40	250	31.7 (1.2)	51.7 (1.2)	8.3 (1.2)	8.3 (1.2)
	500	31.7 (.8)	51.7 (.8)	8.3 (.8)	8.3 (.8)
	1000	31.7 (.6)	51.7 (.6)	8.3 (.6)	8.3 (.6)
Mean		31.7 (.9)	51.7 (.9)	8.3 (.9)	8.3 (.9)

Note. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *TA%* = true accept %; *TR%* = true reject %; *FA%* = false accept %; *FR%* = false reject %.

Table 8
Percent of Selection Successes and Selection Errors

Unit-Weighted Composite

	N	*TA%*	*TR%*	*FA%*	*FR%*
SR = .10	250	6.6 (.7)	86.6 (.7)	3.4 (.7)	3.4 (.7)
	500	6.6 (.5)	86.6 (.5)	3.4 (.5)	3.4 (.5)
	1000	6.6 (.4)	86.6 (.4)	3.4 (.4)	3.4 (.4)
Mean		6.6 (.6)	86.6 (.6)	3.4 (.6)	3.4 (.6)
SR = .20	250	14.5 (1.0)	74.5 (1.0)	5.5 (1.0)	5.5 (1.0)
	500	14.5 (.7)	74.5 (.7)	5.5 (.7)	5.5 (.7)
	1000	14.5 (.5)	74.5 (.5)	5.5 (.5)	5.5 (.5)
Mean		14.5 (.8)	74.5 (.8)	5.5 (.8)	5.5 (.8)
SR = .40	250	32.4 (1.1)	52.4 (1.1)	7.6 (1.1)	7.6 (1.1)
	500	32.4 (.8)	52.4 (.8)	7.6 (.80)	7.6 (.8)
	1000	32.4 (.6)	52.4 (.6)	7.6 (.6)	7.6 (.6)
Mean		32.4 (.9)	52.4 (.9)	7.6 (.9)	7.6 (.9)

Note. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *TA%* = true accept %; *TR%* = true reject %; *FA%* = false accept %; *FR%* = false reject %.

Multiple-hurdle model. Table 9 shows the results for selection accuracy in the predictor for multiple-hurdle selection. Selection errors were more prevalent in the multiple-hurdle model. The results showed that across the number of applicants, mean of 48.1% of the applicants selected based on the observed predictor score actually would not have been selected on at least one of the hurdles had the selection been based on their true predictor score. Although not modeled in the current thesis, selection errors are likely to be more prevalent for multiple-hurdle selection if rejected applicants who would have been selected based on true predictor scores (*false rejects*) are considered.

Table 9
Percent of Selection Successes and Selection Errors

Multiple Hurdles		
N	*TA%*	*FA%*
250	51.5 (9.7)	48.5 (9.7)
500	52.1 (7.1)	47.9 (7.1)
1000	52.0 (4.9)	48.0 (4.9)
Mean	51.9 (7.5)	48.1 (7.5)

Note. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *TA%* = true accept %; *FA%* = false accept %.

Mean Comparisons in the Predictor Scores

Compensatory model. Tables 10 and 11 outline the mean differences in observed predictor scores and true predictor scores between the selected group (*true accepts* and *false accepts*) and the rejected group (*true rejects* and *false rejects*) for the two compensatory model selection conditions.

As was shown in the initial findings, there was a distinct difference between the observed and true mean predictor score for the selected applicants and the rejected applicants. Specifically, due to selection errors in each of the selection decision groups, true mean predictor scores for the selected group were attenuated compared to their corresponding observed mean predictor score, whereas the true mean predictor score for the rejected group increased compared to their corresponding observed mean predictor score. This pattern of results was consistent across the predictor weight conditions.

Table 10
Observed and True Standardized Mean Predictor Composite Scores for Accepted and Rejected Applicants

Regression-Weighted Composite					
	N	*Accepted OS*	*Rejected OS*	*Accepted TS*	*Rejected TS*
SR = .10	250	3.41 (.12)	.65 (.11)	2.75 (.14)	1.31 (.09)
	500	3.42 (.09)	.65 (.08)	2.75 (.10)	1.31 (.06)
	1000	3.43 (.06)	.64 (.05)	2.75 (.07)	1.31 (.04)
	Mean	3.42 (.09)	.65 (.08)	2.75 (.11)	1.31 (.07)
SR = .20	250	2.64 (.07)	.05 (.08)	1.98 (.09)	.72 (.07)
	500	2.65 (.05)	.05 (.06)	1.98 (.06)	.72 (.05)
	1000	2.65 (.04)	.05 (.04)	1.98 (.04)	.72 (.04)
	Mean	2.64 (.06)	.05 (.06)	1.98 (.07)	.72 (.06)
SR = .40	250	1.64 (.06)	-.82 (.07)	.98 (.07)	-.16 (.06)
	500	1.64 (.04)	-.82 (.05)	.98 (.05)	-.16 (.05)
	1000	1.64 (.03)	-.82 (.03)	.98 (.03)	-.16 (.03)
	Mean	1.64 (.04)	-.82 (.05)	.98 (.05)	-.16 (.05)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *Accepted OS* = observed predictor composite score for *observed score accepts*; *Rejected OS* = observed predictor composite score for *observed score rejects*; *Accepted TS* = true predictor composite score for *observed score accepts*; *Rejected TS* = true predictor composite score for *observed score rejects*.

Table 11
Observed and True Standardized Mean Predictor Composite Scores for Accepted and Rejected Applicants

Unit-Weighted Composite	*N*	*Accepted OS*	*Rejected OS*	*Accepted TS*	*Rejected TS*
SR = .10	250	3.39 (.13)	.69 (.10)	2.79 (.14)	1.29 (.09)
	500	3.40 (.08)	.69 (.07)	2.79 (.09)	1.30 (.06)
	1000	3.40 (.06)	.68 (.05)	2.79 (.07)	1.30 (.04)
	Mean	3.40 (.09)	.69 (.08)	2.79 (.11)	1.30 (.07)
SR = .20	250	2.61 (.08)	.09 (.08)	2.01 (.09)	.70 (.07)
	500	2.62 (.05)	.09 (.06)	2.01 (.06)	.70 (.05)
	1000	2.62 (.04)	.09 (.04)	2.01 (.04)	.70 (.04)
	Mean	2.62 (.06)	.09 (.06)	2.01 (.07)	.70 (.06)
SR = .40	250	1.61 (.06)	-.79 (.07)	1.01 (.06)	-.18 (.06)
	500	1.61 (.04)	-.79 (.05)	1.00 (.05)	-.18 (.05)
	1000	1.61 (.03)	-.79 (.03)	1.00 (.03)	-.18 (.03)
	Mean	1.61 (.05)	-.79 (.05)	1.01 (.05)	-.18 (.05)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *Accepted OS* = observed predictor composite score for *observed score accepts*; *Rejected OS* = observed predictor composite score for *observed score rejects*; *Accepted TS* = true predictor composite score for *observed score accepts*; *Rejected TS* = true predictor composite score for *observed score rejects*.

Because the scores were standardized, the differences in means between those selected and rejected can be expressed in terms of standardized mean differences. Across predictor weights and selection ratios, mean differences in the observed predictor scores between the selected group and the rejected group attenuated by more than one standard deviation when the mean comparisons were made based on their corresponding true predictor scores (i.e., change in standardized mean difference was greater than 1). The mean differences attenuated, on average, by 1.33 standard deviation for the regression-weight condition, and by 1.22 standard deviation for the unit-weight condition (see Table 12). Looking into how the scores changed for each selection decision groups, for the regression-weight condition, true mean predictor score for the selected applicants was on average, .66 standard deviation lower than

their corresponding observed mean predictor score. This difference was slightly lower for the unit-weight condition (average standardized mean difference of .61). On the other hand, true mean predictor score for those rejected was on average .66 standard deviation higher than their observed mean predictor score (average standardized mean difference of .61 for unit-weight condition).

Standardized mean differences in terms of Cohen's U_3 showed that the attenuation in the mean predictor score difference between the comparisons based on observed scores to true scores slightly increased as the selection ratio increased (i.e., decrease in mean difference was greater as selection ratio increased). As illustrated in Table 12, the decrease in the percent of the rejected applicants' predictor score exceeded by the mean predictor score of the selected group increased from 7.2% in .10 selection ratio (99.7% - 92.5%) to 12.0% in .40 selection ratio (99.3% - 87.3%) in the regression-weight condition, and 6.5% in .10 selection ratio (99.7% - 93.2%) to 10.9% in .40 selection ratio (99.2% - 88.3%) in the unit-weight condition. Put differently, for the regression-weight condition, the percentile difference between the mean observed predictor score of the rejected group and the mean observed predictor score of the selected group, in terms of the distribution of predictor score of the rejected group, ranged from 49.7% (for a .10 selection ratio) to 49.3% (for a .40 selection ratio). However, the percentile difference between the true mean predictor score of the rejected group and the true mean predictor score of the selected group ranged from 42.5% (for a .10 selection ratio) to 37.3% (for a .40 selection ratio). A similar pattern was found for the unit-weight condition.

Table 12
Observed and True Score Differences in Standardized Mean Predictor Composite Scores between Accepted and Rejected Applicants

Regression-Weighted Composite				
	Accepted OS – Rejected OS	*Accepted TS – Rejected TS*	*OS difference in U_3*	*TS difference in U_3*
SR = .10	2.77	1.44	99.7	92.5
SR = .20	2.59	1.26	99.5	89.6
SR = .40	2.46	1.14	99.3	87.3
Unit-Weighted Composite				
	Accepted OS – Rejected OS	*Accepted TS – Rejected TS*	*OS difference in U_3*	*TS difference in U_3*
SR = .10	2.71	1.49	99.7	93.2
SR = .20	2.53	1.31	99.4	90.5
SR = .40	2.40	1.19	99.2	88.3

Notes. *Accepted OS – Rejected OS* = mean observed predictor score difference between *observed score accepts* and *observed score rejects*; *Accepted TS – Rejected TS* = mean true predictor score difference between *observed score accepts* and *observed score rejects*; *OS difference in U_3* = U_3 for Accepted OS – Rejected OS; *TS difference in U_3* = U_3 for Accepted TS – Rejected TS.

Multiple-hurdle model. Table 13 shows the mean observed and true predictor scores for the selected and the rejected applicants. As was expected, observed mean difference between the selected and the rejected applicants was attenuated when the comparisons were made based on their respective true scores. Note however, that in the multiple-hurdle model, the difference is generally less distinct compared to when selection was made based on compensatory model. As is shown on Table 14, the observed mean predictor score difference 1.92, whereas the true mean predictor score difference was 1.73. Thus, mean predictor score difference between the selection decision groups only attenuated by .19 when the selected and rejected applicants were compared based on their true predictor scores as opposed to their observed predictor scores. Consequently, there was no practical difference between each selection decision group's observed mean predictor score and their corresponding true mean predictor score (standardized mean difference was .14 for

selected applicants; standardized mean difference was .05 for rejected applicants). In line with these results, change in Cohen's U_3 was less than 2% (97.3% - 95.8%).

Table 13
Observed and True Standardized Mean Predictor Composite Scores for Accepted and Rejected Applicants

Multiple-Hurdles				
N	*Accepted OS*	*Rejected OS*	*Accepted TS*	*Rejected TS*
250	1.43 (.15)	-.46 (.04)	1.30 (.16)	-.41 (.05)
500	1.45 (.10)	-.46 (.03)	1.31 (.12)	-.42 (.04)
1000	1.45 (.08)	-.47 (.02)	1.31 (.08)	-.42 (.03)
Mean	1.44 (.11)	-.47 (.03)	1.31 (.12)	-.42 (.04)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *Accepted OS* = observed predictor composite score for *observed score accepts*; *Rejected OS* = observed predictor composite score for *observed score rejects*; *Accepted TS* = true predictor composite score for *observed score accepts*; *Rejected TS* = true predictor composite score for *observed score rejects*.

Table 14
Observed and True Score Differences in Standardized Mean Predictor Composite Scores between Accepted and Rejected Applicants

Multiple-Hurdles			
Accepted OS – Rejected OS	*Accepted TS – Rejected TS*	*OS difference in U_3*	*TS difference in U_3*
1.92	1.73	97.3	95.8

Notes. Accepted OS – Rejected OS = mean observed predictor score difference between *observed score accepts* and *observed score rejects*; *Accepted TS – Rejected TS* = mean true predictor score difference between *observed score accepts* and *observed score rejects*; *OS difference in U_3* = *U_3* for Accepted OS – Rejected OS; *TS difference in U_3* = *U_3* for Accepted TS – Rejected TS.

Mean Comparisons in the Criterion

Compensatory model. Tables 15 to 20 show the results for comparisons in the difference in true mean criterion performance between the selected and the rejected applicants when selection is based on observed predictor score versus true predictor score. These differences are calculated for both unit-weighted and regression-weighted compensatory models.

Results showed that across the predictor weight conditions, the difference in mean criterion performance between *true score accepts* and *true score rejects* was

greater than the mean criterion performance difference between *observed score accepts* and *observed score rejects*.

Table 15
Standardized True Mean Criterion Performance for Observed Score Selection Decision Groups and True Score Selection Decision Groups

Regression-Weighted Composite					
	N	*OS accepts*	*OS rejected*	*TS accepts*	*TS rejected*
SR = .10	250	1.85 (.30)	.89 (.25)	2.30 (.29)	.44 (.25)
	500	1.86 (.21)	.88 (.17)	2.31 (.20)	.43 (.18)
	1000	1.86 (.15)	.88 (.12)	2.30 (.15)	.44 (.12)
	Mean	1.86 (.23)	.88 (.19)	2.30 (.22)	.44 (.19)
SR = .20	250	1.33 (.20)	.48 (.19)	1.77 (.21)	.03 (.18)
	500	1.33 (.15)	.49 (.13)	1.78 (.14)	.04 (.13)
	1000	1.33 (.11)	.48 (.09)	1.78 (.10)	.04 (.09)
	Mean	1.33 (.16)	.48 (.14)	1.78 (.15)	.04 (.14)
SR = .40	250	.66 (.17)	-.11 (.16)	1.11 (.15)	-.55 (.16)
	500	.66 (.12)	-.11 (.11)	1.11 (.11)	-.55 (.11)
	1000	.66 (.12)	-.11 (.11)	1.10 (.08)	-.55 (.07)
	Mean	.66 (.13)	-.11 (.12)	1.10 (.12)	-.55 (.12)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *OS accepts* = true mean criterion performance for *observed score accepts*; *OS rejected* = true mean criterion performance for *observed score rejects*; *TS accepts* = true mean criterion performance for *true score accepts*; *TS rejected* = true mean criterion performance for *true score rejects*.

Table 18
Standardized True Mean Criterion Performance for Observed Score Selection Decision Groups and True Score Selection Decision Groups

Unit-Weighted Composite					
	N	*OS accepts*	*OS rejected*	*TS accepts*	*TS rejected*
SR = .10	250	1.67 (.34)	.78 (.27)	2.03 (.33)	.42 (.29)
	500	1.69 (.23)	.77 (.18)	2.05 (.22)	.41 (.19)
	1000	1.68 (.16)	.77 (.14)	2.03 (.16)	.42 (.14)
	Mean	1.68 (.25)	.77 (.20)	2.04 (.25)	.42 (.21)
SR = .20	250	1.22 (.23)	.41 (.20)	1.56 (.22)	.07 (.20)
	500	1.21 (.16)	.42 (.14)	1.58 (.15)	.06 (.14)
	1000	1.21 (.12)	.41 (.10)	1.56 (.11)	.06 (.11)
	Mean	1.21 (.17)	.42 (.15)	1.57 (.17)	.06 (.16)
SR = .40	250	.61 (.18)	-.11 (.25)	.96 (.18)	-.46 (.17)
	500	.61 (.14)	-.11 (.29)	.96 (.12)	-.46 (.13)
	1000	.61 (.09)	-.11 (.09)	.96 (.09)	-.47 (.08)
	Mean	.61 (.14)	-.11 (.23)	.96 (.14)	-.46 (.13)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *OS accepts* = true mean criterion performance for *observed score accepts*; *OS rejected* = true mean criterion performance for *observed score rejects*; *TS accepts* = true mean criterion performance for *true score accepts*; *TS rejected* = true mean criterion performance for *true score rejects*.

Perhaps more important than this general pattern is illustrating the degree to which the difference in mean criterion performance changes between selection based on a battery of observed predictor scores versus a battery of their corresponding true predictor scores. Tables 16 and 19 show these changes in standardized measures of effect. Across selection ratios for the regression-weight condition, mean criterion performance for *true score selects* was .44 standard deviation higher than the mean criterion performance for *observed score accepts*. Average U_3 across selection ratios was 67%, indicating that in terms of the distribution of the *observed score accepts*, mean criterion performance of the *true score accepts* was 17 percentile points higher than mean criterion performance of the *observed score accepts* (see Table 16). Although the difference was slightly lower for the unit-weight condition, this general

pattern was preserved, with the mean criterion performance for the *true score accepts* being .36 standard deviation higher than the mean criterion performance for the *observed score accepts*. The average U_3 across selection ratios was 64.1%, indicating a 14-percentile difference between the mean of the *true score accepts* and the mean of the *observed score accepts*, in terms of the distribution of the *observed score accepts* (see Table 19).

Table 16
Differences in Standardized True Mean Criterion Performance between True Score Accepts and Observed Score Accepts
Regression-Weight Condition

SR	*TS accepts – OS accepts*	*TS accepts – OS accepts in* U_3
.10	.44	67.0
.20	.45	67.4
.40	.44	67.0

Notes. TS accepts – OS accepts = true mean criterion performance difference between *true score accepts* and *observed score accepts*; *TS accepts – OS accepts in* $U_3 = U_3$ for true mean criterion performance difference between *true score accepts* and *observed score accepts.*

Table 19
Differences in Standardized True Mean Criterion Performance between True Score Accepts and Observed Score Accepts
Unit-Weight Condition

SR	*TS accepts – OS accepts*	*TS accepts – OS accepts in* U_3
.10	.36	64.1
.20	.36	64.1
.40	.35	63.7

Notes. TS accepts – OS accepts = true mean criterion performance difference between *true score accepts* and *observed score accepts*; *TS accepts – OS accepts* in $U_3 = U_3$ for true mean criterion performance difference between *true score accepts* and *observed score accepts.*

Also, mean criterion performance difference between the *true score accepts* and *true score rejects* was greater than the mean criterion performance difference between the *observed score accepts* and *observed score rejects*. For the regression-weighted composite, mean criterion performance difference was greater by .88 to .89

standard deviation across selection ratios for the comparisons between *true score accepts* and *true score rejects* than for the comparisons between *observed score accepts* and *observed score rejects*. For example, for a .10 selection ratio, the standardized mean difference between *true score accepts* and *true score rejects* was 1.86, whereas the difference was .98 between *observed score accepts* and *observed score rejects* (see Table 17). The same pattern of results was found for the unit-weighted composite, but the magnitude of the difference in the standardized means for selection based on true predictor score (i.e., difference between *true score accepts* and *true score rejects*) versus observed predictor score (i.e., difference between *observed score accepts* and *observed score rejects*) was slightly lower compared to the regression-weight condition, with the standardized mean differences ranging from .70 to .72 across selection ratios (see Table 20).

Change in standardized mean difference in terms of U_3 offers another interpretation. For both regression- and unit-weighted conditions, U_3 increased by more than 10% when comparisons were between *true score accepts* and *true score rejects* than when they were between *observed score accepts* and *observed score rejects*. In other words, the percentile difference between the true mean criterion performance of the selected group and the rejected group increased by more than 10% when selection was based on perfectly reliable true predictor scores, compared to when selection was based on unreliable observed predictor scores. As expected, the increase in U_3 was greater as the selection ratio increased, meaning that measurement error variance has a greater effect when an organization can afford to be more selective.

Table 17
Differences in Standardized True Mean Criterion Performance between Observed Score Selection Decision Groups and True Score Selection Decision Groups

Regression-Weight Condition				
SR	*OS accepts – OS rejects*	*TS accepts – TS rejects*	*OS accepts – OS rejects in U_3*	*TS accepts – TS rejects in U_3*
.10	.98	1.86	83.7	96.9
.20	.85	1.74	80.2	95.9
.40	.77	1.65	77.9	95.1

Notes. OS accepts – OS rejected = true mean criterion performance difference between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejected* = true mean criterion performance difference between *true score accepts* and *true score rejects*; *OS accepts – OS rejected in $U_3 = U_3$* for true mean criterion performance difference between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejects in $U_3 = U_3$* for true mean criterion performance difference between *true score accepts* and *true score rejects*.

Table 20
Differences in Standardized True Mean Criterion Performance between Observed Score Selection Decision Groups and True Score Selection Decision Groups

Unit-Weight Condition				
SR	*OS accepts – OS rejects*	*TS accepts – TS rejects*	*OS accepts – OS rejects in U_3*	*TS accepts – TS rejects in U_3*
.10	.91	1.62	81.9	94.7
.20	.79	1.51	78.5	93.5
.40	.72	1.42	76.4	92.2

Notes. OS accepts – OS rejected = true mean criterion performance difference between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejected* = true mean criterion performance difference between *true score accepts* and *true score rejects*; *OS accepts – OS rejected in $U_3 = U_3$* for true mean criterion performance difference between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejected in $U_3 = U_3$* for true mean criterion performance difference between *true score accepts* and *true score rejects*.

Multiple-hurdle model. Similar to the results in the compensatory model condition, mean criterion performance of the *true score accepts* was greater than the mean criterion performance of the *observed score accepts* (see Table 21). However, the difference in criterion performance was only slight. In terms of standardized measures of effect, criterion performance of the *true score accepts* was only .06 standard deviation above criterion performance of the *observed score accepts*.

Consequently, U_3 between the *true score accepts* and *observed score accepts* was only 52.4%, indicating that there was almost a perfect overlap in the distribution of criterion performance between the *true score accepts* and *observed score accepts* (see Table 22).

Table 21
Standardized True Mean Criterion Performance for Observed Score Selection Decision Groups and True Score Selection Decision Groups

Multiple-Hurdles Condition				
N	*OS accepts*	*OS rejected*	*TS accepts*	*TS rejected*
250	.33 (.05)	-.19 (.07)	.39 (.05)	-.22 (.07)
500	.33 (.04)	-.19 (.05)	.39 (.03)	-.23 (.05)
1000	.33 (.03)	-.19 (.04)	.39 (.03)	-.23 (.03)
Mean	.33 (.05)	-.19 (.04)	.39 (.05)	-.23 (.05)

Notes. Values inside the parentheses are the standard deviations for the percentages across 1,000 replications. *OS accepts* = true mean criterion performance for *observed score accepts*; *OS rejected* = true mean criterion performance for *observed score rejects*; *TS accepts* = true mean criterion performance for *true score accepts*; *TS rejected* = true mean criterion performance for *true score rejects*.

Table 22
Differences in Standardized True Mean Criterion Performance between True Score Accepts and Observed Score Accepts

Multiple-Hurdles Condition	
TS accepts – OS accepts	*TS accepts – OS accepts in U_3*
.06	52.4

Notes. TS accepts – OS accepts = difference in mean true criterion performance between *true score accepts* and *observed score selects*; *TS accepts – OS accepts in U_3* = U_3 for difference in mean true criterion performance between *true score accepts* and *observed score accepts*.

Results for the mean criterion performance difference between the selection decision groups was greater when selection was based on true predictor scores compared to when selection was based on observed predictor scores (i.e., difference in mean criterion performance greater between *true score accepts* and *true score rejects* than between *observed score accepts* and *observed score rejects*). However, the change in mean criterion difference was relatively small (standardized mean difference was .62 between *true score accepts* and *true score rejects*, whereas

standardized mean difference was .52 between *observed score accepts* and *observed score rejects*). Distribution overlap in criterion performance between the selected group and the rejected group increased from 69.9% when selection was based on the observed predictor score (i.e., distribution overlap between *observed score accepts* and *observed score rejects*) to 73.2% when selection was based on the true predictor score (i.e., distribution overlap between *true score accepts* and *true score rejects*). In other words, the percentile difference between the mean criterion performance of the selected group and the rejected group increased from 19.9% when selection was based on the observed predictor score to 23.2% when selection was based on the true predictor score (see Table 23).

Table 23
Differences in Standardized True Mean Criterion Performance between Observed Score Selection Decision Groups and True Score Selection Decision Groups Multiple-Hurdles Condition

OS accepts – OS rejects	*TS accepts – TS rejects*	*OS accepts – OS rejects in U_3*	*TS accepts – TS rejects in U_3*
.52	.62	69.9	73.2

Notes. *OS accepts – OS rejects* = difference in true mean criterion performance between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejects* = difference in true mean criterion performance between *true score accepts* and *true score rejects*; *OS accepts – OS rejects in U_3 = U_3* for difference in true mean criterion performance between *observed score accepts* and *observed score rejects*; *TS accepts – TS rejects in U_3 = U_3* for difference in true mean criterion performance between *true score accepts* and *true score rejects*.

Predictor Composite Scores: Unit Weights vs. Regression Weights

Consistent with the initial findings, predictor weights did not have much influence on the percent of selection success and selection error, and their mean predictor scores. However, selected applicants in the regression-weight condition consistently had higher mean criterion performance than the selected applicants in the unit-weight condition for both when selection was based on observed predictor score or true predictor score. For selection based on observed predictor score, standardized

mean difference was .18 for a .10 selection ratio, .12 for a .20 selection ratio, and .05 for a .40 selection ratio. For selection based on true predictor score, standardized mean difference was .26 for a .10 selection ratio, .21 for a .20 selection ratio, and .14 for a .40 selection ratio.

Discussion

Negative consequences of less-than-perfect predictor measures may be exacerbated or reduced based on the features of the data (e.g., validity, reliability of the predictors) or external parameters of the selection situation (e.g., selection ratio, score combination rules) that largely influence the overall worth of the selection instrument. The primary purpose of the current thesis was to illustrate that because of measurement unreliability, what the organization actually gains from selection, in terms of selection accuracy, mean predictor score, and mean criterion performance, is likely lower compared to what the organization could have gained with more reliable predictor battery. In addition to calculating the rates of *selection accuracy* and *selection errors*, I made several mean comparisons in terms of standardized measures of effect that convey the practical gains in predicted performance (as indicated by predictor scores), and criterion performance that organizations can incur through selection based on reliable predictor battery. This work is important because it illustrates the impact that common organizational selection practices have on practical outcomes of the organization.

An important aspect of the study is worth pointing out. As Roth et al. (2011) suggested, careful attention needs to be placed in aligning the population estimates in meta-analytic matrices to be used as input values in a simulation study, and the theoretical level of population at which the conclusions are intended for (e.g., incumbents vs. applicants). Simulations in the current thesis were based on an input

correlation matrix that included recent meta-analytic findings that are more refined. Specifically, the criterion-related validity for cognitive ability was updated to include validity estimates across three-levels of job complexity (high-, medium-, and low-complexity). Also, the correlations were corrected for range restriction. In turn, the correlation values more accurately captured the relationship among the study variables at the construct level, and for the appropriate population (applicant level).

Selection Accuracy

Organizations must make selection decisions based on the applicants' observed scores across predictors. Yet, because each predictor is a less-than-perfectly reliable indicator of the underlying construct it intends to measure, these observed scores will deviate randomly from their corresponding true scores. Because of this, organizations may reject applicants who actually meet the cutoff for selection, or select applicants who do not actually meet the cutoff. Simulation results for compensatory model selection showed that a lower selection ratio was associated with lower rate of selection errors. These results were aligned with the range restriction in the error variance. However, the proportion of selection errors among the selected applicants was slightly higher with greater selectivity (lower selection ratios). In fact, across conditions, results showed that the proportion of selection errors among the selected applicants (*false accepts*) is considerable. Specifically, the mean percentages of *false accepts* among the selected applicants (calculated as the proportion of percentage *false accepts* over percentage selected) were roughly 37% for .10 selection ratio ($\frac{3.69}{10}$), 30% for .20 selection ratio ($\frac{6.01}{20}$), and 21% for .40 selection ratio ($\frac{8.33}{40}$). For the multiple-hurdle model, mean rate of *false accepts* was over 48%. From the organization's perspective, these numbers might be alarming, as falsely

accepted applicants pose a greater problem to its productivity because these applicants are in an immediate position to potentially undermine the organization's performance.

Results of the current thesis illustrated the practical effect that measurement unreliability can have on selection accuracy. Based on the results of the current study, two issues are evident that should be of interest to organizations looking to reduce or minimize the losses due to selection errors. First, it is critical for organizations to evaluate the acceptability of their selection battery in terms of measurement reliability and the level of selection consistency it provides. Second, organizations should actively seek ways to improve reliability of their selection procedures.

Obviously, selection errors are less likely to occur to the extent that the predictor measures are reliable. However, it is not apparent how much reliability is needed to ensure certain level of selection accuracy. Results from the current simulations showed that selection errors were prevalent even when the alpha reliability estimates for each individual predictor variables was higher than the popularly used benchmark of .70, which is probably a very low benchmark especially for personnel selection settings where the personal and legal stakes might be high (Lance et al., 2006).

There are a number of methods that can be used to estimate the level of selection accuracy under given parameters (including the reliability of the measure). However, some of these methods are restrictive, such as applying only to dichotomously scored items, applying only to compensatory models, or lacking direct implications for prediction accuracy and subsequent utility estimation (e.g., Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976). Livingston and Lewis (1995) introduced a classification accuracy estimation method that can be applied to

more variety of conditions where many restrictive assumptions are not necessary. Although the computation process itself is complex, simulations have shown their results to be highly accurate. In this method, observed score X_i is transformed to a scale ranging from 0 to 1 by subtracting the minimum score of the measurement scale from X_i, and dividing this difference by the range of the measurement scale ($X_{max} - X_{min}$). Then, based on the distribution of observed proportional scores, distribution of the true proportional scores is estimated based on a method developed by Lord (1965; see also Hanson, 1991, pp. 3-9). Classification accuracy is then estimated by the rate of agreement between the type of classification (selected vs. rejected) based on the test taker's observed scores and their corresponding true scores. Livingston and Lewis (1995) applied this method to estimate the classification accuracy statistics in seven different types of tests that differed from each other in the content, format, and statistical characteristics (multiple-choice test used in the licensing of elementary school teachers, Advanced Placement Program, and a test consisting of holistically scored essays or problems for teachers seeking alternate-route certification). Each test was divided into two half-tests, and the applicants' selection consistency in each of the two half-tests was described as the *actual classification accuracy rate* of the test. Livingston and Lewis applied their method to data from each half-tests. The results showed that for all seven tests, selection accuracy estimates were within .02 of the actual classification rates.

It is strongly recommended that organizations analyze the expected values and variability of selection success (or error) rates based on reasonable estimates of their own selection conditions prior to the application of a predictor battery. Simulation scenarios may be the best way to conduct the analysis, though as mentioned there may be limited situations where direct psychometric formulas may apply. Such practice

has the potential to improve selection and save costs that are involved in replacing erroneously hired applicants.

Even with these a priori measures to minimize measurement error variance or estimate the extent to which selection errors will occur under a given condition, it should be noted that in practice, selection accuracy may not be as high as anticipated because measurement reliability is conditional upon characteristics that are specific to the sample (e.g., ability, motivation, interaction between the items and the applicants) that the organization does not have full control over (Hambleton, Swaminathan, & Rogers, 1991).

There are several sources of error in the measurement process, such as test-taker or rater characteristics or motivations, measurement length, and the general quality or complexity of the items. As briefly mentioned earlier, applying generalizability theory can provide valuable information regarding the influence that such factors have on measurement error (Hambleton & Slater, 1997). Several studies have shown that generalizability theory can be applied to credentialing tests that are often used as a component of a selection battery, and that their results are dependable (e.g., Brennan, 1992). More importantly, these studies have shown that they can provide practical answers regarding potential ways to improve measurement reliability. For example, Brennan and Johnson (1995) applied generalizability theory to examine the variance in measurement error that number of raters and measurement length provide. Results showed that longer measures assessed by single raters are less affected by error variance than shorter measures assessed by multiple raters. Thus, it is recommended that researchers and practitioners might consider how more complex forms of measurement error might influence the accuracy of selection and prediction,

and then develop practical recommendations that then can make the selection procedure more accurate.

Results also showed that organizations could reduce selection errors that disadvantage them by being more selective. Doing so restricts the range of error variance among the selected applicants, thus the true scores of the applicants selected based on restrictive selection are likely to be higher and therefore more acceptable, even if they deviate (are lower than) the observed score. This notion is in line with Feldt, Steffen, and Gupta's (1985) findings. With all five SEM estimation methods the authors used in their study, they found evidence that the level of standard error is conditional upon the raw score distribution of the measure. Specifically, standard error peaked at the middle of the score distribution (SEM was greater), and gradually declined as the scores reached either extremes of the distribution. Despite the psychometric advantages however, it may not be practical or possible to be more selective (e.g., there are certain number of slots that must be filled; cost associated with recruitment). Improved measurement on the predictor and criterion sides of the equation, combined with improvements in recruiting and training, might be more viable recourses.

Mean Comparisons in the Predictor

As noted, measurement error variance can meaningfully alter the applicants' observed predictor scores from their corresponding true scores. Because selection is based on the unreliable observed scores of the predictors, selection errors occur where mean true scores of the falsely hired applicants are attenuated compared to their respective observed scores, and mean true scores of the falsely rejected applicants are higher compared to their respective observed scores. To illustrate the effect that measurement unreliability has on mean predictor score, I compared the mean

difference between the selected applicants (*true accepts* and *false accepts*) and the rejected applicants (*true rejects* and *false rejects*) for both observed and true scores.

Across the two predictor-weight conditions in the compensatory model, mean observed predictor score for the selected applicants was greater compared to the mean observed predictor score for the rejected applicants by more than two standard deviations regardless of the number of applicants or the selection ratio. Their corresponding true score difference however, was attenuated by more than one standard deviation for each condition. For example, under the regression-weight condition, with a .10 selection ratio, the standardized mean difference in the observed mean predictor score between selected applicants and rejected applicants was 2.77. However, difference in their corresponding true score was only 1.44. This general pattern was also found in the multiple-hurdle model, but the change in mean predictor score difference from observed score to true score was much smaller compared to the results for compensatory model condition.

Organizations make their selection decisions based on predictor scores that are thought to measure constructs related to job performance. Given that there is a linear relationship between the predictor battery and the criterion of interest, top-down selection on a predictor composite should lead to improved performance on the criterion. Thus, given the same selection parameters, a predictor battery that provides a more distinct difference between the selected group and the rejected group in terms of mean scores is more valuable for the purpose of achieving a higher level of performance from selection than a predictor battery that does not. The results of the current simulations showed that the observed mean score difference between the selection decision groups does not convey the whole story. Rather, because of measurement error, the distinction that is actually made between the selected

applicants and the rejected applicants in terms of mean predictor score is actually much smaller than what the organization thought to have gained from using the predictor battery.

Mean Comparisons in the Criterion

Perhaps of more immediate interest to organizations, I examined how measurement unreliability affects the desired outcomes of selection in terms of mean difference in true criterion performance when selection was based on observed predictor scores versus true predictor scores.

As expected, there was a greater difference in mean criterion performance between *true score accepts* and *true score rejects* than between *observed score accepts* and *observed score rejects*. Simulations were conducted not to verify this obvious expectation, but to determine the amount of these differences under realistic conditions. For compensatory model selection, on average, the difference between *true score accepts* and *true score rejects* was .88 standard deviation greater than the difference between *observed score accepts* and *observed score rejects* for the regression-weight condition, and .72 standard deviation for the unit-weight condition.

Also, there was a distinct difference in true mean criterion performance between the *observed score accepts* and the *true score accepts*. Specifically, in the regression-weight condition, criterion performance for the *true score accepts* was .44 standard deviations higher than the criterion performance for the *observed score accepts*. The difference was .36 standard deviations for the corresponding groups in the unit-weight condition. Although these standardized mean differences translate to small-to-medium effect based on Cohen's rule of thumb where .20 is a "small" effect and .50 is a "medium" effect (Cohen, 1992), they may have important practical impact depending on the situation. For example, an increase in the dollar value

associated with even a small increase in criterion performance may be steep depending on the other parameters such as SD_y (Schmidt et al., 1979). In addition, I only compared the mean differences in criterion performance for individuals in isolation, which does not capture the added benefit that "star players" provide by single-handedly lifting the performance of coworkers – or the cost of not hiring them. Of course, there is also the benefit of not hiring "bad apples" who spoil the performance of an entire group – or the cost of erroneously selecting them due to measurement error variance. To the extent that these effects are true, then the benefits of reliable predictor scores is likely to be greater than those expressed in these simulation results. Likewise, the costs of measurement error variance – the unrealized positive and realized negative effects – are likely greater than those shown here.

In contrast to the compensatory model conditions, in the multiple-hurdle condition, difference in mean criterion performance between *true score accepts* and *true score rejects* was only slightly greater compared to the difference in mean criterion performance between *observed score accepts* and *observed score rejects*. Specifically, criterion score difference between the selected group and the rejected group increased by only .10 standard deviation when selection was based on true predictor scores rather than the observed predictor scores. Similarly, the mean criterion performance for the *true score accepts* was only .06 standard deviation higher than the mean criterion performance for the *observed score accepts*. This can be attributed again to the fact that selection is not optimized at each step of the multiple-hurdle selection process. Again, the general pattern of the results is in line with the expectations, but the actual magnitude of these differences has implications

for the usefulness of reliable predictor scores and compensatory vs. multiple-hurdle selection methods.

Predictor Composite Scores: Regression Weights vs. Unit Weights

Current results are in accord with historical findings (Ghiselli et al., 1981; Guilliksen, 1950; Wainer, 1976; Wilks, 1938), that regression-weighted and unit-weighted composite scores were very similar (high correlations between them for each condition), leading to a small effect size difference in criterion performance between applicants selected based on the regression-weighted vs. unit-weighted predictor scores. Unit-weights are generally preferred to avoid the potential for capitalization on chance when using regression-based scores, especially if composite selection score must be generated based on small sample sizes (Bobko, Roth, & Buster, 2007).

Predictor Score Combination Models

There are fundamental differences in how selection is conducted in compensatory model and multiple-hurdle model. Thus, the general choice of the selection model has important effect on the applicants who are selected, and their predicted or actual performance (Chester, 2003). Via simulation, the current thesis demonstrated distinct differences between the two selection methods. As indicated earlier, longer composite on the compensatory selection model provide more reliable selection results than selection on each constituent measurements in the predictor battery, as is done in the multiple-hurdle model. Despite this important advantage, multiple-hurdle model is a more popular method of selection because of practical limitations associated with compensatory model selection (e.g., cost). Any differences in the simulation results thus provide basis for comparisons between a more reliable vs. a more popular selection method.

Note that even where only 7% of the applicant pool was selected, the mean predictor score and actual criterion performance for selected applicants under the multiple-hurdle model were generally lower compared with that for the compensatory model. This is to be expected because multiple-hurdle selection creates direct and incidental range restriction that reduces variance in both the predictor and the criterion (Guion, 1998; Sackett, Laczo, & Arvey, 2002). Specifically, selection on the cognitive ability test in the initial hurdle creates direct range restriction on that variable, but that creates incidental range restriction not only on the criterion, but also on the second and third hurdles where direct selection (range restriction) is yet to occur. Applicants might have scored above the direct cutoff on the second and third hurdles, but were eliminated because they did not pass the first hurdle. This leads to mean scores for each predictor that are generally lower than in compensatory model condition, and therefore mean predictor score will be lower as well. In short, selection is not optimized for the cutoff on each predictor variable in multiple-hurdle selection.

It is also worth noting that in the multiple-hurdle model, the change in the degree of observed mean versus true mean predictor score composite difference between the selected group and the rejected group is smaller compared with the corresponding change in the compensatory model condition. This is also the case for the difference in true mean criterion performance between the *observed score accepts* and the *true score accepts*. These results can also be attributed to the non-optimal range restriction effects in multiple-hurdle selection that was mentioned previously. Although multiple-hurdle models are more popular, they are less efficient in producing desired outcomes.

Multiple factors influence the effectiveness of predictor scores used in personnel selection. However, these influences can get ignored or lost in translation when communicating how they affect practical outcomes of personnel selection strategies. The current thesis examined how much influence measurement error variance has on the quality of selection, illustrated in terms *selection accuracy*, or the consistency between observed scores leading to the same selection decision as true scores, and their implications for actual criterion performance of the applicants in each group defined by selection accuracy (i.e., true accepts, true rejects, false accepts, false rejects).

Limitations and Future Directions

Several limitations of the current research should be noted as indicators of where future research on how selection and prediction accuracy contribute to utility. First, although our correlations are based on current meta-analytic estimates of reliability, predictor intercorrelations, and validity, these estimates in specific situations may deviate in meaningful ways that are not due to sampling error. Second, the correlations involving structured interview and biodata may be especially suspect because these measures are methods and not constructs; they therefore can be quite different depending on the selection context or the construct(s) that the methods were designed to capture (Arthur & Villado, 2008). For the purposes of the present simulation, parameter estimates were intended to be reasonable but not only to illustrate the general principles of the effect of measurement unreliability on selection accuracy and performance utility; however, future research could vary the number of predictors and patterns of correlation to investigate specific situations and/or more general properties.

Third, the current simulations were based on the unidimensional job performance assumption on the grounds that multidimensional job performance criterion has often not been justified empirically when performance ratings across dimensions are highly intercorrelated (Van Iddekinge & Ployhart, 2008; Viswesvaran et al., 2005). Note that the use of overall job performance does not necessarily contradict multidimensional theories of job performance (e.g., Campbell et al., 1996). Rather, overall job performance may be thought of as a combination of distinct-but-correlated dimensions that might be literally or subjectively averaged across different dimensions (Rotundo & Sackett, 2002). In fact, several studies have shown that overall supervisory performance ratings reflect both task and contextual job performance factors (e.g., Borman, White, & Dorsey, 1995; Johnson, 2001; Kiker & Motowidlo, 1999; MacKenzie, Podsakoff, & Fetter, 1991; Orr, Sackett, & Mercer, 1989; Werner, 1994). Thus, it is advised that even though real world data on job performance might be unidimensional, performance measurement still needs to involve multidimensional theories of job performance.

Fourth, future research might examine multiple criteria simultaneously, as other simulations have demonstrated that consideration of multiple criteria (and variability in weights given to each component) can greatly influence composite validity. For example, Murphy and Shiarella (1997) showed that by varying the weights for the predictors (cognitive ability and conscientiousness) and the criteria (task and contextual performance), and the standard deviations of the two criterion performance dimensions, the central 95% of the distribution of obtained values across the weightings of predictors and criteria ranged from .20 to .78. Results indicated that depending on the weights placed on the predictors and the criteria (or theoretically, how the organization chooses to define performance), the validity of a predictor

battery might substantially improve or decrease. Thus, future research on selection and prediction accuracy needs to consider a future where data from performance measures support multidimensionality, for one because they are more reliable than the dismal average level of reliability of .52 found in Viswesvaran and Ones' (2000) meta-analysis.

Fifth, future studies should examine how organizations might change their selection practices given these findings. Organizations could decide to raise or lower their selection ratio. However, the number of needed applicants and the minimally acceptable standard for satisfactory performance is not likely to change. Thus, organizations might consider how recruiting and training contribute to selection utility formulas by improving the yield of qualified applicants pre-hire and by endowing applicants with qualifications post-hire, respectively.

Sixth, the current simulations rely on a CTT approach to reliability estimation. Future research could extend the simulations by identifying sources of systematic error variance and use of generalizability to model this error and estimate reliability in a more complex and realistic manner.

Concluding Comments

The current thesis extended results found in the traditional utility tables by capturing a more complete representation of selection accuracy, beyond what is provided in the Taylor-Russell model. In addition, I compared mean predicted performance (MPP) from a predictor composite of observed scores versus MPP from a similar composite of true scores. Results indicate specific practical gains in performance that could be realized by improving the reliability of predictor measures (i.e., reducing both selection errors and prediction errors). This is more detailed and informative than what a reliability coefficient alone would tell you.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435-442. doi: 10.1037/0021-9010.93.2.435

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Ben-David, M. F. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher, 22,* 120-130.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478-494. doi: 10.1037/0021-9010.74.3.478

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 562-589. doi: 10.1111/j.1744-6570.1999.tb00172.x

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80,* 168-177. doi: 10.1037/0021-9010.80.1.168

Boudreau, J. W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology, 36,* 551-576. doi: 10.1111/j.1744-6570.1983.tb02235.x

Boudreau, J. W., Sturman, M. C., & Judge, T. A. (1994). Utility analysis: What are the black boxes, and do they affect decisions? In N. Anderson, P. Herriot (Eds.), *Assessment and selection in organizations: Methods and practice for recruitment and appraisal* (pp. 77-96). New York: Wiley.

Bradlow, E. T., & Wainer, H. (1998). Some statistical and logical considerations when rescoring tests. *Statistica Sinica, 8,* 713-728.

Brennan, R. L. (1992). NCME instructional module: Generalizability theory. *Educational Measurement: Issues and Practice*, *11*, 27-34. doi: 10.1111/j.1745-3992.1992.tb000260.x

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24,* 339-353.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38,* 295-317. doi: 10.1111/j.1745-3984.2001.tb01129.x

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*, 9-12. doi: 10.1111/j.1745.3992.1995.tb00882.x

Bretz, R. D., Milkovich, G. T., & Read, W. The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, *18*, 321-352. doi: 10.1177/014920639201800206

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37,* 64-76. doi: 10.1037/h0061548

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2,* 171-185. doi: 10.1111/j.1744-6570.1949.tb01397.x

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (p. 258-299). San Francisco: Jossey-Bass.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interview: Raising the psychometric properties of the employment interview. *Personnel Psychology*, *41*, 25-42. doi: 10.1111/j.1744-6570.1988.tb00630.x

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology*, *71*, 672-678. doi: 10.1037/0021-9010.71.4.672

Cascio, W. F. (1980). Responding to the demands for accountability: A critical analysis of three utility models. *Organizational Behavior and Human Performance, 25,* 32-45.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cascio, W. F., & Boudreau, J. W. (2008). *Investing in people: Financial impact of human resource initiatives*. Upper Saddle River, NJ: Pearson.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233-264. doi: 10.1207/s15327043hup0404_1

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22,*32-14. doi: 10.1111/j.1745-3992.2003.tb00126.x

Crano, W. D., & Brewer, M. B. (1973). *Principles of research in social psychology*. New York: McGraw-Hill.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi: 10.1007/BF02310555

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi: 10.1037/0033-2909.112.1.155

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*, 565-579. doi: 10.1037/0021-9010.80.5.565

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104. doi: 10.1037/0021-9010.78.1.98

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). *Personnel Psychology*, *53*, 325-351. doi: 10.1111/j.1744-6570.2000.tb00204.x

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297-300. doi: 10.1037/0021-9010.75.3.297

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12,* 1-16.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory for generalizability of scores and profiles*. New York: Wiley.

Curtis, E. W., & Alf, E. F. (1969). Validity, predictive efficiency, and practical significance of selection tests. *Journal of Applied Psychology, 53,* 327-337. doi: 10.1037/h0027852

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106. doi: 10.1037/h0037613

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674. doi: 10.1126/science.2648573

Dean, M. A. (2004). An assessment of biodata predictive ability across multiple performance criteria. *Applied H.R.M. Research*, *9*, 1-12.

De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology, 91,* 523-537. doi: 10.1037/0021-9010.91.3.523

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*, 1380-1393. doi: 10.1037/0021-9010.92.5.1380

Dipboye, R. L., Gaugler, B. B., Hayes, T. L., & Parker, D. (2001). The validity of unstructured panel interviews: More than meets the eye? *Journal of Business and Psychology, 16,* 35-49. doi: 10.1023/A:1007883620663

Distefano, M. K., & Pryer, M. W. (1987). Evaluation of selected interview data in improving the predictive validity of a verbal ability test with psychiatric aide trainees. *Educational and Psychological Measurement, 47,* 189-192. doi: 10.1177/0013164487471027

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics, 35,* 1–27. doi: 10.3102/1076998609346969

Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's (1988) model and the General Aptitude Test Battery. *Journal of Applied Psychology, 86,* 60-79. doi: 10.1037/0021-9010.86.1.60

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9*, 351-361. doi: 10.1177/014662168500900402

Gatewood, R. D., & Field, H. S. (2004). *Human resource selection* (5th ed.). Mason, OH: South-Western Thompson Learning.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for behavioral sciences*. San Francisco, CA: W. H. Freeman.

Gresham, F. M. (2003). Establishing the technical adequacy of functional behavioral assessment: Conceptual and measurement challenges. *Behavioral Disorders, 28,* 282-298.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures:

The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293-323. doi: 10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19-30. doi: 10.1037/1040-3590.12.1.19

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hakstian, A. R., Woolley, R. M., Woolsey, L. K., & Kryger, B. R. (1991). Management selection by multiple-domain assessment: II. Utility to the organization. *Educational and Psychological Measurement, 51,* 899-911. doi: 10.1177/001316449105100410

Hambleton, R., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19-38. doi: 10.1207/s15324818ame1001_2

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Research Rep. No. 91-5). Iowa City, IA: American College Testing.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164. doi: 10.1177/014662168500900204

Hattrup, K., O'Connell, M. S., & Labrador, J. R. (2005). Incremental validity of locus of control after controlling for cognitive ability and conscientiousness. *Journal of Business and Psychology, 19*, 461-481. doi: 10.1007/s10869-005-4519-1

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection, *1*, 333-342. doi: 10.1111/j.1754-9434.2008.00058.x

Hills, J. R. (1971). Use of measurement in selection and placement. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 680-732). Washington, DC: American Council on Education.

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92,* 555-566. doi: 10.1037/0021-9010.92.2.555

Hoffman, C. C., & Thornton, G. C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*, 455-470. doi: 10.1111/j.1744-6570.1997.tb00916.x

Hom, P. W., & Griffeth, R. W. (1991). Structural equations modeling test of a turnover theory: Cross-sectional and longitudinal analyses. *Journal of Applied Psychology, 76,* 350-366. doi: 10.1037/0021-9010.76.3.350

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362. doi: 10.1016/0001-8791(86)90013-8

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. *Journal of Applied Psychology, 62,* 245-260. doi: 10.1037/0021-9010.62.3.245

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869-879. doi: 10.1037/0021-9010.85.6.869

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253-264. doi: 10.1111/j.1745-3984.1976.tb.00016.x

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86,* 984-996. doi: 10.1037/0021-9010.86.5.984

Kaiser, R. B., Adorno, A. J., Williams, K. B., & Binning, J. F. (1996, April). *A field study of gender and race effects in a structured panel interview.* Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*, 179-182. doi: 10.1007/BF02289635

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17,* 221-240. doi: 10.1207/s15324818ame1703_1

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin, 127,* 673-697. doi: 10.1037/0033-2909.127.5.673

Kiker, D. S., & Motowidlo, S. J. (1999). Main and interaction effects of task and contextual performance on supervisory reward decisions. *Journal of Applied Psychology, 84,* 602-609. doi: 10.1037/0021-9010.84.4.602

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9,* 202-220. doi: 10.1177/1094428105284919

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72-107. doi: 10.1037/0033-2909.87.1.72

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12,* 165-200. doi: 10.1177/1094428107302900

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179-197. doi: 10.1111/j.1745-3984.1995.tb00462.x

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, *16*, 247-260. doi: 10.1111/j.1745-3984.1979.tb00106.x

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika, 27,* 19-30.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239-270. doi: 10.1007/BF02289490

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Mosley.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of salespersons' performance. *Organizational Behavior and Human Decision Processes, 50,* 123-150. doi: 10.1016/0749-5978(91)90037-T

Mattson, J. T. (2003). The effects of alternative reports of human resource development results on managerial support. *Human Resource Development Quarterly, 14,* 127-151. doi: 10.1002/hrdq.1056

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599-616. doi: 10.1037/0021-9010.79.4.599

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational and Behavioral Statistics, 12,* 292-293. doi: 10.3102/10769986012003282

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93-115. doi: 10.1037/1082-989X.9.1.93

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*, *89*, 158-164. doi: 10.1037/0021-9010.89.1.158

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873-900. doi: 10.1111/j.1744-6570.2000.tb02421.x

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50,* 823-854. doi: 10.1111/j.1744-6570.1997.tb01484.x

Myors, B. (1993). A Taylor-Russell/Naylor-Shine utility calculator. *Behavioral Research Methods Instrument & Computers, 25,* 483-484. doi: 10.3758/BF03204549

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology, 3,* 33-42.

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Orr, J. M., Sackett, P. R., & Mercer, M. (1989). The role of prescribed and nonprescribed behaviors in estimating the dollar value of performance. *Journal of Applied Psychology, 74,* 34-40. doi: 10.1037/0021-9010.74.1.34

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dillema: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Selection*, *61*, 153-172. doi: 10.1111/j.1744-6570.2008.00109.x

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment, 13,* 304-315. doi: 10.1111/j.1468-2389.2005.00327.x

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology*, *49*, 85-102. doi: 10.1111/j.1744-6570.1996.tb01792.x

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173-184. doi: 10.1177/01466216970212006

Ree, M. J., & Carretta, T. R. (1994). Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Assessment, 54,* 459-463. doi: 10.1177/0013164494054002020

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1,* 407-420. doi: 10.1177/109442819814003

Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *International Journal of Selection and Assessment, 7,* 157-169.

Roth, P. L., Bobko, P., Switzer, F. S., & Dean, M. A. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology, 54,* 591-617. doi: 10.1111/j.1744-6570.2001.tb002224.x

Roth, P. L., & Campion, J. E. (1992). An analysis of the predictive power of the panel interview and pre-employment tests. *Journal of Occupational and Organizational Psychology, 65,* 51-60. doi: 10.1111/j.2044-8325.1992.tb00483.x

Roth, P. L., Switzer, F. S., Van Iddekinge, C. H., & Oh, I. S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology, 64,* 899-935. doi: 10.1111/j.1744-6570.2011.01231.x

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175-184. doi: 10.1037/0021-9010.75.2.175

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policty-capturing approach. *Journal of Applied Psychology, 87,* 66-80. doi: 10.1037/0021-9010.87.1.66

Rozeboom, W. W. (1989). The reliability of a linear composite of nonequivalent subtests. *Applied Psychological Measurement, 13,* 277-283. doi: 10.1177/014662168901300307

Rudner, L. (2001). Computing the expected proportion of misclassified examinees. *Practical Assessment, Research and Evaluation, 7.* Retrieved August 16, 2010 from http:// PAREonline.net/getvn.asp?v1⁄47&n1⁄414

Rusbult, C. E., & Farrell, D. (1983). A longitudinal test of the investment model: The impact of job satisfaction, job commitment, and turnover of variations in rewards, costs, alternatives, and investments. *Journal of Applied Psychology, 68,* 429-438. doi: 10.1037/0021-9010.68.3.429

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413-428. doi: 10.1037/0033-2909.88.2.413

Sackett, P. R., De Corte, W., & Lievens, F. (2010). Decision aids for addressing the validity-adverse impact trade-off. In J. L. Outtz (Ed.), *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection* (pp. 453-472). New York, NY: Routledge/Taylor & Francis Group.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, *55*, 807-825. doi: 10.1111/j.1744-6570.2002.tb00130.x

Sackett, P. R., & Roth, L. (1991). A Monte Carlo examination of banding and rank order methods of test use in personnel selection. *Human Performance, 4,* 279-295. doi: 10.1207/s15327043hup0404_3

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 549-572. doi: 10.1111/j.1744-6570.1996.tb01584.x

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*, 302-318. doi: 10.1037/0003-066X.56.4.302

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929-954. doi: 10.1037/0003-066x.49.11.929

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003a). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, *56*, 573-605. doi: 10.1111/j.1744-6570.2003.tb00751.x

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003b). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88,* 1068-1081. doi: 10.1037/0021-9010.88.6.1068

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199-223. doi: 10.1037/1082-989x.1.2.199

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274. doi: 10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology, 64,* 609-626. doi: 10.1037/0021-9010.64.6.609

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U. S. park ranger for three modes of test use. *Journal of Applied Psychology, 69,* 490-497. doi: 10.1037/0021-9010.69.3.490

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53,* 901-912. doi: 10.1111/j.1744-6570.2000.tb02422.x

Schmitt, N., Pulakos, E. D., Nason, E., & Whitney, D. J. (1996). Likability and similarity as potential sources of predictor-related criterion bias in validation research. *Organizational Behavior and Human Decision Processes*, *68*, 272-286. doi: 10.1006/obhd.1996.0105

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719-730. doi: 10.1037/0021-9010.82.5.719

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265-276. doi: 10.1111/j.1745-3984.1976.tb00017.x

Swaminathan, H. Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267. doi: 10.1111/j.1745-3984.1974.tb00998.x

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23,* 565-578. doi: 10.1037/h0057059

U.S. Equal Opportunity Employment Commision, U. S. Civil Service Commission, U.S. Department of Labor, U.S. Department of Justice, (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38295-38309.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871-925. doi: 10.1111/j.1744-6570.2008.00133.x

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 224-235. doi: 10.1177/00131640021970475

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557-574. doi: 10.1037/0021-9010.81.5.557

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108-131. doi: 10.1037/0021-9010.90.1.108

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217. doi: 10.1037/0033-2909.83.2.213

Werner, J. M. (1994). Dimensions that make a difference: Examining the impact of in-role and extrarole behaviors on supervisory ratings. *Journal of Applied Psychology, 79,* 98-107. doi: 10.1037/0021-9010.79.1.98

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40. doi: 10.1007/bf02287917

Youngblood, S. A., Mobley, W. H., Meglino, B. M. (1983). A longitudinal analysis of the turnover process. *Journal of Applied Psychology, 68,* 507-516. doi: 10.1037/0021-9010.68.3.50